Customizing Machine Learning to Combat Cyberattacks

Abstract

Cyber criminals are on the warpath. Not only have cyberattacks increased in numbers and frequency, but their growing sophistication and success rate also point to the inability of commercial cyber security solutions to detect and counter them. The available commercial security products have not been able to take advantage of the progress made in the areas of machine learning and artificial intelligence. Yet, combining custom-built ML solutions with mature commercial cyber security products can help enterprises win the war against cyber criminals. This paper recommends a three-pronged approach to achieve this: develop robust ML-based algorithms hardened against potential subversion and degradation by attackers; employ active learning to develop ML-based algorithms using limited training data; and develop customized ML-based algorithms trained using client-specific data in the client-specific environment.

Introduction

Cyber security breaches have been rising steadily, despite the growing number of state-of-the-art security products and platforms for endpoint protection. According to a study by Ponemon Institute, the global average cost of a breach was more than $3.5 million in 2017.[1] Over the next two years, the study estimates, there is a 27% likelihood of a recurring material data breach. In addition to direct losses to data, security breaches cause loss of reputation and brand image, and a decline in consumer confidence. Behind this never-ending spiral of security breaches is the inability of commercial off-the-shelf (COTS) security products to adapt quickly and promptly to advances in malware development, which have not been able to leverage the latest innovations in machine learning (ML) and artificial intelligence (AI) in the same way that malware has.

Rapid Advances in Malware Development

Hackers are using more and more advanced weapons every day--polymorphic and metamorphic malware, non-malware attacks, and ML--to develop new types of attack and to subvert detection software. Experts predict that ML will be used to improve the success and effectiveness of the attacks; therefore, the solutions, too, must intelligently incorporate the same technology.

Verifying that a new file is malicious can be complex and time consuming, and often lags behind the malware evolution, rendering corporations vulnerable to serious damages. The Cisco 2017 Annual Cybersecurity Report found that many malware families took less than 24 hours' time to evolve (TTE).[2] A worrying new development is using non-malware for malicious attacks. These are capable of gaining control of computers without the user having to download any malicious files; the attacker uses existing software, approved applications, and authorized protocols to carry out malicious activities. According to a survey by Carbon Black, a leading cyber security company, "nearly two-thirds (64%) of security researchers said they have seen an increase in non-malware attacks since the beginning of 2016."[3]

Of late, a growing number of attacks seeks to subvert ML-based detection. There are two main ways to attack an ML-based system: manipulating the learning system to allow a specific attack, and degrading the system to make it unusable.

One example of such attacks is the "Boiling Frog Poisoning," where the attacker slowly inject malicious training data periodically before the attack as the ML-based anomaly detector is continuously retrained.

From Signature-based to Machine Learning Detection

Traditional approaches in malware detection are signature-based for static analysis and rule-based for dynamic analysis. Signature-based detection uses digital signatures of binary objects and matches them with known signatures of previously discovered malicious objects. This approach can neither detect any new objects nor effectively detect polymorphic and metamorphic malware or non-malware attacks.

Rule-based detection uses predefined patterns in software behavior to detect malicious software. Here, too, a new previously unseen behavioral pattern will go initially undetected until some harm is done. "Smart" malware can detect the environment it enters and change its behavior to evade detection.

Most security products, traditionally, have used rule-based software to detect malware, network intrusions, etc., which is no longer sustainable. Signature-based detection has been considered inadequate since the early 2000s.[4] At the same time, the cybersecurity community largely considers ML-based solutions at a nascent stage; 70% of security researchers told the Carbon Black survey that attackers could bypass ML-driven technologies.[5]

While it is obvious how hackers can easily avoid signature-based detection, it is also easy to see how they can learn to avoid ML-based detection. The wide availability of COTS security products is a double-edged sword. A well-organized and well-funded cybercrime enterprise can easily access these products and services to harden their attack tools. There is some evidence that hackers hijack commercial antivirus programs to bypass all security protection.[6] Since these programs are considered trusted, hackers can get complete control over the data on various devices.

Applying machine learning to counter cyberattacks is preferable to rule- or signature-based approaches. There are two types of machine learning that have been used in the field of cyber security: classification using supervised learning, and anomaly detection using semi-supervised learning. The basic concept of

supervised learning as applied in static analysis consists of presenting a computer algorithm with multiple samples of benign and malicious software. The algorithm using advanced ML techniques learn different patterns and is capable of classifying never-seen-before objects as benign or malicious.

The same approach can be applied in dynamic analysis. In this case, a computer algorithm is presented with multiple examples of benign and malicious behavior. Successful algorithms are capable of generalizing the differences and classifying new previously unseen behavior instances as benign or malicious. The downside of the supervised classification approach is that the classifiers may not be able to detect truly new and conceptually different attacks. However, it is important to recognize that these limitations are not as severe as those of the rule-based/signature-based approaches. The patterns found by machine are much more complex than any human can detect. Another difference is that these patterns can be continually modified based on new known malicious patterns by "simply" feeding the new data into the classifiers.

Semi-supervised anomaly detection is what can be used to detect these conceptually different malware attacks. A computer algorithm is trained on multiple examples of benign software or benign software behavior. The algorithm creates a probabilistic model, which is used to estimate the probability of any unseen object or behavior to be malicious. A mature solution should comprise both classification and anomaly detection algorithms working together.

Custom-tailored Solutions Based on Machine Learning

Developing a successful security solution to safeguard organizations effectively against sophisticated cyberattacks will need a three-pronged strategy:

- Develop robust ML-based algorithms hardened against potential subversion and degradation by attackers.

  An ML-based system can be manipulated to allow a specific attack; it can even be degraded to make it completely unusable. Identification of potential vulnerabilities of the ML algorithms can be done proactively by simulating different types of attacks. Robust statistical and ML methods developed to counteract simulated attacks will be resistant to subversion and degradation.

- Employ active learning to develop ML-based algorithms, using limited training data.

  One of the main reasons for only partial success of the ML-based solutions is that the data needed for training is complex and not readily available. Active learning, which is a special case of semi-supervised machine learning, can address that challenge. An active learning algorithm is designed to obtain an ongoing feedback from the user on the accuracy of its prediction. Security experts will be engaged to identity false positives, which will be used to improve the existing algorithms using limited training datasets.

- Develop customized ML-based algorithms trained using client-specific data in the client-specific environment.

  ML-based solutions must be custom-tailored to meet the needs of individual customers. Each customer faces specific attacks, based on their business, size, location, etc. A custom-tailored solution will provide customers with a unique cybersecurity protection solution, tailored exclusively to their needs. This will also protect the solution from the reach of the worldwide hacker.

Conclusion

The solution we advocate does not make existing COTS products completely irrelevant. Ultimately, the ideal solution will be a hybrid of the existing COTS products and custom-tailored ML-based solutions. Only a hybrid solution can meet the increasing need for reliable, secure and consistently high performance against cyberattacks.

References

1. Ponemon Institute, 2017 Cost of Cyber Crime Study, October1, 2017, https://www.ponemon.org/library/2017-cost-of-cyber-crime-study, accessed June 28, 2018
2. Cisco, 2017 Annual Cybersecurity Report, http://www.cisco.com/c/dam/m/digital/1198689/Cisco_2017_ACR_PDF.pdf, accessed June 28, 2018
3. Carbon Black, Beyond the Hype, 2017, https://www.carbonblack.com/wp-content/uploads/2017/03/Carbon_Black_Research_Report_NonMalwareAttacks_ArtificialIntelligence_MachineLearning_BeyondtheHype.pdf, accessed June 28, 2018
4. Edward Hurley, Virus management: Never a Dull Moment, TechTarget, https://searchsecurity.techtarget.com/tip/Virus-management-Never-a-dull-moment
5. Carbon Black, Ibid
6. Catalyn Cimpanu, New Attack Uses Microsoft's Application Verifier to Hijack Antivirus Software, March 21, 2017, https://www.bleepingcomputer.com/news/security/new-attack-uses-microsofts-application-verifier-to-hijack-antivirus-software/, accessed June 28, 2018

About The Authors

Satish Thiagarajan - Global Head – Cyber Security Practice
Over 25 years of experience across industries and IT involving consulting, business analysis, process re-engineering, concurrent multi-project delivery management in Application support, Infrastructure management, Enterprise Security and Testing, Currently Head Cyber Security and Risk Management Unit – Responsible for service strategy, GTM, delivery strategy, competency development and partner development.

David Makovoz
David Makovoz is a Lead Data Scientist developing Machine Learning solutions for the TCS Cyber Security Unit. He has over 20 years of a unique blend of experience spanning data science, algorithm and software development, distributed computing, mathematical and computer modeling, complex system architecture and design. Throughout his career he has lead numerous innovative, leading edge initiatives across a wide array of complex and diverse domains including data mining, predictive analytics, machine learning and deep learning, image, speech and natural language processing, agent-based, system dynamics and discrete event modeling, social network analysis. His strengths are the ability to successfully translate academic theories and concepts to business solutions and a focus on developing practical solutions to challenging technical problems. David received his Ph.D. in Nuclear Physics from the University of Washington.

Contact

Visit the Cyber Security page on www.tcs.com

Email: cyber.security@tcs.com

Subscribe to TCS White Papers

TCS.com RSS: http://www.tcs.com/rss_feeds/Pages/feed.aspx?f=w
Feedburner: http://feeds2.feedburner.com/tcswhitepapers

About Tata Consultancy Services Ltd (TCS)

Tata Consultancy Services is an IT services, consulting and business solutions organization that delivers real results to global business, ensuring a level of certainty no other firm can match. TCS offers a consulting-led, integrated portfolio of IT and IT-enabled, infrastructure, engineering and assurance services. This is delivered through its unique Global Network Delivery Model™, recognized as the benchmark of excellence in software development. A part of the Tata Group, India's largest industrial conglomerate, TCS has a global footprint and is listed on the National Stock Exchange and Bombay Stock Exchange in India.

For more information, visit us at www.tcs.com

TCS Design Services | M | 08 | 18