Machine Learning for Cyber Security using Big Data Analytics

R.Vijaya Lakshmi

Assistant Professor

Department of Computer Science & Engineering

Alliance University, Bangalore

Corresponding author's email id: *vijaya.lakshmi@alliance.edu.in**

DOI: *http://doi.org/ 10.5281/zenodo.3362228*

Abstract

Machine Learningis an Approach to AI that uses a system that is capable of learning from experience. It is intended not only for AI goals (e.g., copying human behavior) but it can also reduce the efforts and/or time spent for both simple and difficult tasks like stock price prediction. In other words, ML is a system that can recognize patterns by using examples rather than by programming them. Big data analytics in security involves the ability to gather massive amounts of digital information to analyze, picture and draw insights that can make it possible to predict and stop cyber attacks. Along with security technologies, it gives us stronger cyber defense posture. They allow organizations to recognize patterns of activity that represent network threats. In this paper, I emphasis on how Big Data be able to progress information security best practices. I am trying to apply machine learning procedures in cyber security using big data Analytics.

Keywords: *Machine Learning Big Data, Cyber Security, Artificial Intelligence, Deep Learning.*

INTRODUCTION

Machine learning means solving certain tasks with the use of an approach and worse, now hackers are able to use machine learning to carry out all their nefarious endeavors. Fortunately, machine learning can aid in solving the most areas where machine learning is thriving. There will always be a man trying to find weaknesses in systems or ML algorithms and to bypass security mechanisms. If your system learns constantly, makes decisions based on data rather than algorithms, and change its behavior, it's Machine Learning. Many papers cover machine learning for cyber security and the ability to protect us from cyber attacks. Still, it's important to scrutinize how actually Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning (DL) can help in cyber security right now, and what this hype is all about.

Unfortunately, machine learning will never be a silver bullet for cyber security compared to image recognition or natural language processing, two common tasks including regression, prediction, and classification. In the era of extremely large amount of data and cyber security talent shortage, ML seems to be an only solution [1].This paper is an introduction written to give practical technical understanding of the current advances and future directions of ML research applied to cyber security and Big Data Analytics.

Big Data Analytics

The term big data is devoted to massive amount information that is been stored and spread in a computer system[3].Big Data is differentiated from traditional technology in 3ways:

1. The amount of data - Size: the volume of datasets is a critical factor, that is, how much amount of data that is been generated.

2. The rate of data generation and transmission (Velocity) -Complexity: the structure, behavior and permutations of datasets in critical factor.

3. The types of structured and unstructured data (Variety) - Technologies: tools and techniques that are been used to process a sizable or complex datasets is a crucial factor.

A Cyber Security Company's Big Data Analytics Approach

We began producing antivirus and encryption products nearly 30 years ago, now helps secure the networks used by 100 million people in 150 countries and 100,000 businesses using big data analytics. Today, big data analytics is integral to Sophos' daily malware detection in multiple use cases:

1. Malware research and analysis. Malware is becoming more evasive and pervasive. Sophos analyzes the characteristics of suspicious files and report the analysis outcome.

2. Macro trend analysis. Sophos analysts also analyze the data for macro trends

of malware movements to better understand and anticipate the direction of the threat landscape.

3. Measuring detection performance. Evaluating statistics on the performance of malware detection to understand which protection technology is as long as most value.

How Big Data Helps Avoid Cyber security Threats

Cybercrime instances seem to be refinement like bucks. According to security software maker Malware bytes, its users reported 1 billion malware-based incidents from June to November 2016.That was two years ago. Just picture this figure in 2018. Malware attacks have become more sophisticated and more difficult to detect and fight. Keeping precious business data protected against malware and hacking is one of the biggest challenges facing modern businesses. These never-ending cyber security threats make it extremely difficult to sustain business performance and growth.

Big Data: a savior?

More or less many say Big Data is a threat; others declare it a protector. Big Data can store large amounts of data and help analysts examine, observe, and detect irregularities within a network. That makes Big Data analytics an appealing idea to help escape cyber crimes. The security-related information available from Big Data reduces the time required to detect and resolve an issue, allowing cyber analysts to predict and avoid the possibilities of intrusion and invasion. Insights from Big Data analytics tools can be used to detect cyber security threats, including malware/ransom ware attacks, compromised and weak devices, and malicious insider programs. This is where Big Data analytics looks most promising in improving cyber security. Big Data doesn’t provide rock-solid security due to poor mining and the absence of experts who know how to use analytics trends to fix gaps.

INTELLIGENT RISK MANAGEMENT

To improve your cyber security efforts, your tools must be backed by intelligent risk-management insights that Big Data experts can easily interpret. The key purpose of using these automation tools should be to make the data available to analysts more easily and quickly. This approach will allow our experts to source, categorize, and handle security threats without delay.

Threat visualization

Big Data analytics programs can help us foresee the class and intensity of cyber security threats. We can weigh the complexity of a possible attack by evaluating data sources and patterns. These tools also allow us to use current and historical data to get statistical understandings of which trends are acceptable and which are not.

Predictive models

Intelligent Big Data analytics enables experts to build a predictive model that can issue an alert as soon as it sees an entry point for a cyber security attack. Machine learning and artificial intelligence can play a major role in developing such a mechanism. Analytics-based solutions enable you to predict and gear up for possible events in your process.

Stay secure and ahead of hackers with penetration testing

Infrastructure penetration testing will give us insight for our business database and process and help keep hackers at bay. Penetration testing is a simulated malware attack against our computer systems and network to check for exploitable vulnerabilities. It is like a mock-drill exercise to check the capabilities of our process and existing analytics solutions. Penetration testing has become an essential step to protect IT infrastructure and business data.

Penetration testing involves five stages:

1. Planning and reconnaissance
2. Scanning
3. Gaining access
4. Maintaining access
5. Analysis and Web application firewall (WAF) configuration

The results shown by a penetration test exercise can be used to enhance the fortification of a process by improving WAF security policies. Sometimes vulnerabilities in an infrastructure are right in front of the analysts and property owners and still manage to go unnoticed. Operating systems, services and application flaws, improper configurations, and risky end-user behavior are some of the most common places where cyber security vulnerabilities exist.

Big Data analytics solutions, backed by machine learning and artificial intelligence, give hope that businesses and processes can be kept secure in the face of a cyber security breach and hacking. Employing the power of Big Data, you can improve your data-management techniques and cyber threat-detection mechanisms.

Monitoring and improving your approach can bulletproof your business. Periodic penetration tests can help ensure that your analytics program is working perfectly and efficiently.

MACHINE LEARNING TASKS AND CYBER SECURITY

Let's see the examples of different methods that can be used to solve machine learning tasks and how they are related to cyber security tasks.

Regression

Regression (or prediction) is simple. The knowledge about the existing data is utilized to have an idea of the new data. Take an instance of house prices prediction. In cyber security, it can be applied to fraud detection. The features (e.g., the total amount of suspicious transaction, location, etc.) determine a probability of fraudulent actions. As for technical aspects of regression, all methods can be divided into two large categories: machine learning and deep learning. The same is used for other tasks. For each task, there are the examples of ML and DL methods.

Machine learning for regression

Below is a short list of machine learning methods (having their own advantages and disadvantages) that can be used for regression tasks.

- Liner regression
- Polynomial regression
- Ridge regression
- Decision trees
- SVR (Support Vector Regression)
- Random forest

DEEP LEARNING FOR REGRESSION

For regression tasks, the following deep learning models can be used:

- Artificial Neural Network (ANN)
- Recurrent Neural Network (RNN)
- Neural Turing Machines (NTM)
- Differentiable Neural Computer (DNC)

CLASSIFICATION

Classification is also straightforward. Imagine you have two piles of pictures classified by type (e.g., dogs and cats). In terms of cyber security, a spam filter separating spams from other messages can serve as an example. Spam filters are probably the first ML approach applied to Cyber security tasks. The supervised learning approach his usually used for classification where examples of certain groups are known. All classes should be defined in the beginning.

Challenges to Cyber Security & How Big Data Analytics Can Help

Due to the complication of IT networks have grown-up, the inventiveness and sophistication of cyber security threats and attacks has grown just as quickly. Some sobering stats:

- During June and November of 2016, nearly one billion malware-based incidences occurred

- The estimated cost of cybercrime is up to $1 billion

- 99 percent of computers are vulnerable to cyber attacks

The definitions show that cyber security field refers mostly to machine learning (not to AI). And a large part of the tasks are not human-related. Particular methods based on data we have. Most of tasks are subclasses of the most common ones, which are described below.

- Regression (or prediction)—a task of predicting the next value based on the previous values.

- Classification—a task of separating things into different categories.

- Clustering—similar to classification but the classes are unknown, grouping things by their similarity.

- Association rule learning (or recommendation)—a task of recommending something based on the previous experience.

- Dimensionality reduction—or generalization, a task of searching common and most important features in multiple examples.

- Generative models—a task of creating something based on the previous knowledge of the distribution.

There are different approaches in addition to these tasks. You can use only one approach for some tasks, but there can be multiple approaches for other tasks. As malware attacks increase in volume and complexity, it's becoming more difficult for traditional analytic tooling and infrastructure to keep up thanks to:

Data volume: For example, each day at Sophos Labs, over 300,000 new potentially malicious files that require analysis are reported.

Scalability: SQL-based tooling plus infrastructure doesn’t measure well and is costly to maintain.

CONCLUSION

There are more areas left. I have outlined the basics. On the one hand, machine learning is definitely not a silver-bullet solution if we want to protect our systems. There are many issues with interpretability particularly for deep learning algorithms, but humans also cannot interpret their own decisions. On the other hand, with the growing amount of data and decreasing number of experts it is better to start right now. As the complexity of IT networks has grown, the inventiveness and sophistication of cyber security threats and attacks has grown just as quickly, we are leading the largest source of cyber risk data and analytics with millions of companies modeled for cyber risk in our database, benchmarking individual companies and their third parties. ML is an only remedy. It works now and will be mandatory soon.

REFERENCES

I. Big Data and Specific Analysis Methods for Insurance FraudDetection Ana-Ramona BOLOGA, Razvan BOLOGA, AlexandraFLOREA University of Economic Studies, Bucharest, Romania.

II. Big Data Cyber security Analytics Research Report – Ponemon Institute© Research Report Date: August 2016.

III. Richard A. Derrig, ”Insurance Fraud”, The Journal of Risk and Insurance”, 2002, Vol.69, No.3,271-287.

IV. Bresfelean, Vasile Paul, Mihaela Bresfelean, Nicolae Ghisoiu, andCalin-Adrian Comes. 2007. "Data Mining Clustering Techniques in Academia." In ICEIS (2), pp. 407-410.

V. Bresfelean, V. P., Bresfelean, M., Ghisoiu, N., & Comes, C. A.2008. Determining students’ academic failure profile founded ondata mining methods. In Information Technology Interfaces, IEEE,pp. 317-322.

VI. CLOUD SECURITY ALLIANCE Big Data Analytics for Security Intelligence.

VII. Bryant, Katz, & Lazowska, 2008.

VIII. Big Data Analytics for Detection of Frauds in Matrimonial Websites Vemula Geeta et al | International Journal of Computer Science Engineering and Technology (IJCSET) | March 2015 | Vol 5, Issue 3, 57-61 http://www.ey.com/Publication/vwLUAssets/EY_Big_data:_changing_the_way_businesses_operate/%24FILE/EY-Insights-on-GRCBig.

IX. Baccianella S., Esuli A., Sebastiani F., "Senti Word Net 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining" in LREC, pp-2200–2204, 2010.

X. Dave K., Lawrence S., Pennock D. M., "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews", in Proceedings of the 12th International Conference on World Wide Web, ACM, pp-519–528, 2003.

XI. Liu B., "Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies", San Rafael, Calif, Morgan & Claypool, 5(1): pp-1–167, 2012.

XII. Ohana B., Tierney B., "Sentiment Classification of Reviews using Senti Word Net Machine Learning", in 9th IT&T Conference, Dublin Institute of Technology, Dublin, Ireland, pp-13, 2009.

Cite this Article As

R.Vijaya Lakshmi (2019) **Machine Learning for Cyber Security using Big Data Analytics** Journal of Artificial Intelligence, Machine Learning and Soft Computing, 4(2) 1- 8

http://doi.org/10.5281/zenodo.3362228