DATA MINING IN SECURITY

Rafael Marques Braga
School of Electrical and Computer Engineering
University of Miami
Miami, FL
rxm403@miami.edu

Sharadh Ramesh
School of Computer Science
University of Miami
Miami, FL
sxr323@miami.edu

ABSTRACT

With the increase in the amount of data that is available in the world currently, it becomes all the more important to be able to have robust techniques to protect this information. There are two types of security that we cover in this paper. The first is cyber security which detects threats such as malware attacks and phishing attacks. The other type is national security where the National Security Agency uses data to find patterns on people which could indicate suspicious behavior which could be a potential national threat. In the first case, data mining techniques such as clustering, and classification are used extensively. For instance, in a phishing attack, clustering is used to select the features that a particular email possesses and then classification is applied to classify an e-mail as fake or real. When it comes to national security threats, metadata tagging is used heavily to detect what kind of information a particular individual is accessing or browsing. This gives the data miner an idea of the kind of behavioral patterns a particular person displays which could be indicative of their personality.

This paper shows how the results of such data mining techniques have been very positive and has helped prevent significant number of frauds and terrorist attacks. It also goes over its accuracy and how the cost of employing such methods is definitely justified when compared with the losses that have been mitigated or prevented.

1. INTRODUCTION

Data mining is the process of examining large databases to discover new patterns, predict future trends and extract unknown knowledge from these data sets in an automatic or semi-automatic manner [1]. There is a massive number of applications that data mining can be applied to as huge amounts of data is created on a daily basis. Security is one of the main applications for data mining as security becomes increasingly important in everyone's lives.

In the security side, this paper will focus on how data mining can be used in both cyber security and national security. However, there are multiple other applications for data mining to be included when it comes to security. These can include but are not limited to: the security of the private sector of companies, prevention of fraud in insurance, banking fraud, money laundering recognition and detection of crimes into public places among others.

On the cyber security side, data mining is hugely beneficial as it is one of the 4 main methods used today to monitor malwares and the safety of a network along with scanning, activity monitoring and integrity checking [2]. However, all the other methods can also include data mining algorithms for them to be accomplished correctly and in an efficient manner. Data mining is also used in cyber security for intrusion detection and fraud detection, among others.

Today, national security also heavily depends on data mining. As millions of gigabytes of data are created, we are dependent on data mining tools for pre-processing, tagging, and analyzing the data using quick methods. These tools can include Naïve Bayes method, probabilistic modeling, decision trees and deep learning, among others.

2. DATA MINING & CYBER SECURITY

2.1. Cyber Security

Cyber security is the area protecting our digital network from cyber terrorism, cyber-attacks, and intrusions. These can include unauthorized access to a network, denial of service attacks as well as insider threats [3]. Today, the largest corporations spend billions of dollars to make sure their data is secure when stored digitally. Data mining is in the core of that security to prevent and recognize multiple attacks and malware from happening.

In cyber security there are hundreds of attacks which usually present some common behavior for the system. Using the aid from data mining, these behaviors can be quickly noticed by systems and prevent attacks from happening. Some of the examples of cyber attacks that can be noticed from data mining are DoS attacks, Heartbleed attacks, Infiltration Attacks among others [4].

DoS attacks, or Denial-of-Service is a cyber attack where the attacker makes a target computer or network temporarily

inaccessible due to excessive traffic in the network. This can also be a DDoS or a Distributed Denial-of-Service attack, where the attacker uses multiple systems to bring the targeted computer, service or network down. By using multiple computers to bring down the network, the attacker will flood all the bandwidth and resources of the target with requests, making the target inaccessible to others due to lack of resources [5].

Heartbleed attacks allows attackers to get information leaked by the target's memory. Infiltration attacks are attacks that usually happen from the inside of the network where a vulnerability is found (from any common software) and the attacker explores these vulnerabilities to gain access to the system using a back door [4]. This will open the targeted user to multiple other attacks where the attacker can even gain full access to the system.

2.2 Using Data Mining to recognize Cyber Attacks

Today, multiple companies and a lot of research goes into making the digital world safer and ensuring that large corporations are secure from cyber-attacks, or at least have a very low response time to such attack. Since most of cyber attacks follow a pattern or have a certain behavior, these can be easily flagged when compared to historical data from older attacks.

Data Mining is heavily used for malware and intrusion detection using both classification and clustering techniques. Classification algorithms such as Naïve Bayes approach, artificial neural networks, decision trees and support vector machines are some of the most common algorithms used to detect malwares and attacks. Malware is a malicious piece of code that is designed to damage, disrupt, or gain unauthorized access to a computer system [6].

Clustering is also used to group malware samples with similar characteristics and behavior. Clustering is a type of unsupervised machine learning algorithm which creates groups of similar objects based on their attributes.

One example of clustering being used is to detect intrusions to systems and/or networks. To detect intrusion detection, features extracted from programs are analyzed to detect host-based attacks and the network traffic is analyzed to detect network-based attacks.

2.2.1 Malware detection using data mining and machine learning

As mentioned, data mining techniques are a big part of how malware in networks and devices are detected today. Machine learning methods can take in hidden examples from data sets of both malware and benign code [7]. Then, the classification algorithms as the ones mentioned above will detect patterns from the data given to the algorithm, with the final goal to classify given pieces of data as malware or benign code.

A very common way to figure out the patterns from the data and be able to use it to classify algorithms are decision trees and neural network models.

Decision trees are a decision support tool which makes sequential, hierarchical decisions about outcome variables, based on the predictor data in order to find a solution or make a decision [8]. It is a way to display an algorithm that only contains conditional control statements.

2.2.2 Decision Trees to recognize cyber attacks

There are multiple studies that investigate how to detect cyber-attacks using data mining and machine learning methods such as the ones mentioned. A study used a Hidden Markov Model, artificial neural networks and pattern recognition system based cyber security techniques to detect cyber-attacks [9]. In another paper by Rahman, Al-Saggaf and Zia [4], a framework was built to show an example of how a data mining framework and decision trees can be used to predict and prevent cyber-attacks in cyber security.

In the same study, it first extracted the patterns from historical data of cyberattacks using a J48 decision tree algorithm from WEKA and then built a prediction model to predict future cyber-attacks [4]. This framework followed the diagram illustrated in Figure 1.

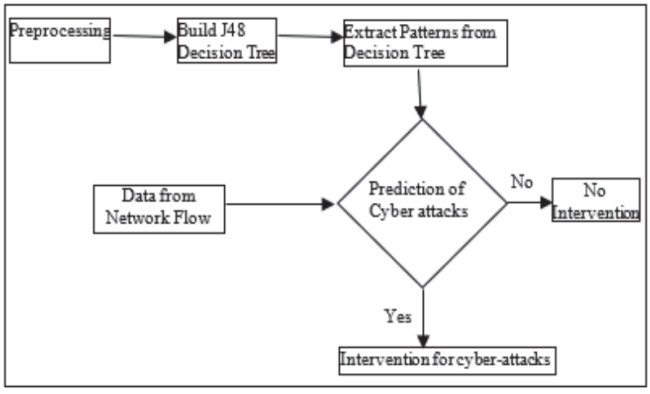

Figure 1 [4]

To evaluate the performance of the prediction model illustrated above, the model used the publicly available cyber security datasets provided by the Canadian Institute of Cybersecurity. The prediction model correctly detected cyber-attacks such as DDoS, PortScan, Bot, Brute force, SQL Injection and Heartbleed with patterns that can be

observed from experimental results. The overall accuracy of the prediction model to detect cyber-attacks on all datasets is around 99%. The accuracy and the extracted patterns for cyber-attacks can be used to predict any future cyber-attacks [4].

A decision tree is a good classification algorithm to be used in this scenario since decision trees are easy to be implemented, easy to process, commonly used to predict and produces logic rules that are easy to understand. However, it can bring some disadvantages like the need to have a good data set for the decision tree to be reliable otherwise some overfitting may occur.

The experimental results of the prediction model used in this study highlight the superiority of data mining models and its capacity to detect future cyber-attacks with a simple implementation of a classification algorithm. This model shows how data mining is beneficial to the cyber security world and how it has a massive positive impact towards the digital safety of individual networks, the privacy of businesses and the security of governments.

2.2.3 Naïve Bayes classification used in phishing attacks

Phishing is a mechanism by which the phisher sends a fake e-mail, which appears to be sent from a legitimate organization or person, asking the recipient to input personal information such as bank details, username, password, credit card details and so on. There are multiple studies which have been conducted to show that data mining techniques can be used to detect phishing attacks. One study by Prasanta Kumar Sahoo proposed an algorithm to analyze emails [10]. The algorithm is given in the diagram below:

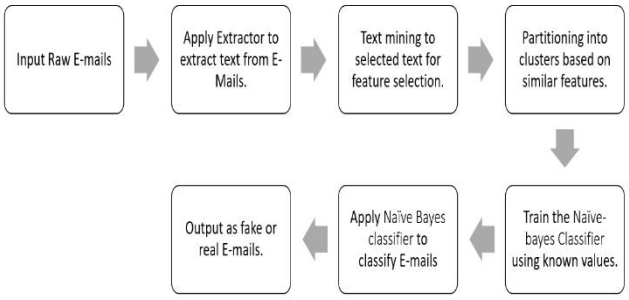

Figure 2

The system first reads the raw e-mails using a java application. In the second stage the content of the e-mail is extracted from the input e-mails. In the third stage text mining is applied to the content to perform feature selection such as personal, family, official, finance and legal issues etc. Subsequently, clustering is applied based on similar features. In the fifth stage, a Naïve Bayes classifier model is used to classify an e-mail as real or fake. If an e-mail is fake, it alerts the user to not provide their personal information or click on any links or attachments that may be provided in the e-mail [10].

The results of this algorithm were analyzed based on the false positive rate (FPR) and the false negative rate (FNR). There may be two types of situations where a false positive can occur:

1. Identifying a non-login page as a login page incorrectly. The overall FPR was 0.68% for this case.
2. Identifying a genuine E-mail as fake E-mail or phishing E-mail incorrectly. The overall FPR was 0.64% for this case [10].

There are also two situations in which a false negative can occur:

1. Not being able to identify a login page. The overall false negative rate was 0.33% for this case.
2. Not being able to identify a phishing site. The overall false negative rate was 0% for this case [10].

3. DATA MINING & NATIONAL SECURITY

Data mining is also used heavily in national security by agencies such as the National Security Agency. They have been used extensively to not only prevent terrorist attacks, bombings, and potential threats to the nation but have also helped in uncovering the perpetrators of the attacks. There have been many case studies that found data mining to be useful. Two famous examples of these are the Tsarnaev brothers who were responsible for the Boston pressure-cooker bombings and the capture of Khalid Ouazzani who was caught funding the Al Qaeda and was suspected of bombing the New York Stock Exchange.

One of the biggest challenges the NSA faces in using data mining is the quantum of data that they must sift through to find patterns in terrorism. There is a huge difficulty in trying to classify this data to find meaningful patterns which can lead them to find suspicious activity. As a result, the methods and the tools used by the NSA are highly sophisticated. One of the techniques that are used in data mining by the NSA is metadata tagging [11]. Metadata is essentially data about data. A label is then placed on data, called a tag, that enables algorithms to identify connections. Tagging data is necessary for data mining because it allows the user to classify the information so that it can be searched. Since communications of US citizens and permanent residents cannot be accessed without a warrant, tagging helps the user to access the information without opening the contents [11]. Even though this is a gray area, using metadata on a tag is not expressly outside the law and consequently, analysts can use it to identify suspicious behavior legally. Tagging is a way in which NSA

overcomes the problem of having huge amounts of data to classify.

These tags are necessary to link different kinds of data such as video, documents, and phone records. For example, data mining could call attention to a suspect on a watch list who downloads terrorist propaganda, visits bomb-making websites, and buys a pressure cooker. This pattern would match the behavior of the Tsarnaev brothers, who perpetrated the Boston Marathon bombings [11].

One source of intelligence and surveillance known as PRISM involves the collection of digital photos, stored data, file transfers, emails, chats, videos, and video conferencing from various internet companies. It is said this platform was used to catch Khalid Ouazzani, a U.S. citizen who was suspected to blow up the New York Stock Exchange. Ouazzani was in contact with an extremist in Yemen, which came to the surface with the NSA. PRISM identified Ouazzani as a possible conspirator and provided information to the FBI. Even though the plot to bomb the New York Stock Exchange did not manifest, Ouazzani pleaded guilty to laundering money to the Al-Qaida [11]. This is a good example of data mining efforts producing results beyond what is expected.

4. CONCLUSION

Data Mining is one of the most powerful automated tools for security in general. With a vast amount of public, historical, and organized data, data mining can be easily implemented and provide huge benefits to businesses and companies when it comes to cyber security. Data collection and data sorting are some of the most difficult tasks when it comes to security due to its high amount of untagged data. With sophisticated tagging mechanisms in place, the NSA and other such security agencies have helped to prevent hundreds of thousands of catastrophic events including terrorist attacks, hackings, and frauds. Consequently, even though there is a cost attached to mining vast amounts of data, it is justified by the amount of money and lives saved along with preventing internet frauds and thefts.

5. REFERENCES

1. [1] Bharati, M. & Ramageri, Bharati. (2010). Data mining techniques and applications. Indian Journal of Computer Science and Engineering. 1.

2. [2] S. User, “Using data mining techniques in cyber security solutions,” Apriorit, [Online]. Available: https://www.apriorit.com/dev-blog/527-data-mining-cyber-security.

3. [3] B. Thuraisingham, "Data mining and cyber security," Third International Conference on Quality Software, 2003. Proceedings., 2003, pp. 2-, doi: 10.1109/QSIC.2003.1319078.

4. [4] M. A. Rahman, Y. Al-Saggaf and T. Zia, "A Data Mining Framework to Predict Cyber Attack for Cyber Security," 2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2020, pp. 207-212, doi: 10.1109/ICIEA48937.2020.9248225.

5. [5] M. A. Saleh and A. Abdul Manaf, "Optimal specifications for a protective framework against HTTP-based DoS and DDoS attacks," 2014 International Symposium on Biometrics and Security Technologies (ISBAST), 2014, pp. 263-267, doi: 10.1109/ISBAST.2014.7013132.

6. [6] “What is malware? - definition and examples,” Cisco, [Online]. Available: https://www.cisco.com/c/en_in/products/security/advanced-malware-protection/what-is-malware.html.

7. [7] Souri, A., Hosseini, R. A state-of-the-art survey of malware detection approaches using data mining techniques. Hum. Cent. Comput. Inf. Sci. 8, 3 (2018). https://doi.org/10.1186/s13673-018-0125-x

8. [8] T. Plapinger, “What is a decision tree?,” Medium, [Online]. Available: https://towardsdatascience.com/what-is-a-decision-tree-22975f00f3e1.

9. [9] M. Kalech, "Cyber-attack detection in SCADA systems using temporal pattern recognition techniques", Computers & Security, vol. 84, pp. 225-238, 2019

10. [10] P. K. Sahoo, "Data mining a way to solve Phishing Attacks," 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), 2018, pp. 1-5, doi: 10.1109/ICCTCT.2018.8550910.

11. [11] Pappallardo, J. (2017, November 14). *NSA data MINING: How it works*. Popular Mechanics. Retrieved September 30, 2021, from https://www.popularmechanics.com/military/a9465/nsa-data-mining-how-it-works-15910146/