DATA MINING IN CYBERSECURITY

Yordan Shterev
"Vasil Levski" National Military University, Veliko Tarnovo, Bulgaria
jshterev@abv.bg

АНАЛИЗ НА ДАННИ В КИБЕРСИГУРНОСТТА

Abstract

The article presents, summarizes and develops technological concepts of cyber security based on machine learning, data processing and analysis. The seven steps of cyber attacks, use of key techniques of data analysis in cyber defense, application of data processing and analysis in information security and some tools used are presented. Applying data processing and analysis as well as extracting dependencies from them in the field of cyber security has some advantages and disadvantages wich are indicated. A common scheme for data analysis is introduced. Main types of penetration of cyberattacks are involved. Cyber attack models abuse detection and anomaly detection are presented too. Three types of anomalies the timing, the number, and the pattern(s) are revealed and examples are pointed.

Keywords: *Cyber Security; Machine Learning; Data Processing and Analysis.*

INTRODUCTION

The development of information and communication technology hardware led to the accumulation of data of different formats, and subsequently to large amounts of such data. This has led to the development of fields such as data processing and analysis and the accompanying interdisciplinary fields - algorithmization, databases, machine learning, statistics, graphical representation of data and others. On the other hand, human nature has led to the creation of viruses, trojans, worms and other malicious software in cyberspace. In addition, for the same reason, various types of attacks in cyberspace have also evolved targeting software, data of various nature, hardware with or without the use of specified malware. From here arose the need to create and develop information security and cyber defense. According to the European Commission, information security is the protection of networks and information systems against human errors, natural disasters, technical malfunctions or malicious attacks [1]. Cyber defense focused on malicious attacks is a complex of preventions and capabilities to protect and actively reaction against cyber attacks and hybrid impacts on communication and information systems and defense and armed forces management systems, as well as on country management systems in a state of emergency, military situation or a state of war and on strategic objects that are relevant to national security [2].

That why, the use and application of the data processing and analysis in information security (cybersecurity) is naturally required [3], [4], [5], [6], [7], [8]. Combining data mining with cyber security provides an opportunity to determine the characteristics of cyber attacks, improve attack detection processes and their countermeasures. In order to obtain valuable knowledge about cyber defense, analysis and data extraction methods from statistics, machine learning, neural networks, database systems and others are used [4], [9], [10].

The purpose of this paper is to introduce, generalize and develop technological concepts of cyber security based on machine learning, the data processing and analysis and dependency extraction.

Steps of a Cyber Attack

In order to consider the place of data processing and analysis in cyber security, it is necessary to clearly distinguish the steps through which it is necessary to go through to realize a *cyber attack*. It is an attempt to destroy, disclose, alter, disable, steal or gain unauthorized access to/or unauthorized use of an information asset [2]. CKC (*Cyber Kill Chain*) has been used for many years by the United States Department of Defense on the battlefield, and now in cyber defense as well. A cyberattack consists of seven sequential steps [11], [12], [13]:

1. Reconnaissance

During this phase, the attacker identifies the target and explores vulnerabilities within the network. It can grab email addresses, user IDs, physical coordinates, login credentials, software applications, operating systems characteristics. This data can be useful in phishing or fraudulent attacks. Honeypots for hosts and networks (systems to hijack cyberattacks) work at this step.

Types of intelligence security:

- Tactical – indicators of compromise, unusual traffic, changes in behavior, failed login attempts;
- Operational – characteristics of the attacker's behavior, the time of the attack, the intention and the behavior;
- Strategic – cybersecurity trends, threat changes.

Data analysis and dependency extraction are typically used only for the first stages of threat intelligence: data discovery and structuring. A cybersecurity expert must then manually review the exposed data and decide what action to take on it.

2. Weaponization

During this phase, the attacker creates an attack vector through which he can exploit a vulnerability discovered in the first step. Attack vectors are virus, worm, remote access malware, ransomware and others. An attacker can also create back doors so that they can continue accessing the system if their original point of entry is identified and closed by network administrators.

3. Delivery

In this step, the attacker launches the attack. The specific steps taken will depend on the type of attack being planned. For example, an attacker can perform a phishing attack by sending email attachments or a malicious link. Thus, the user is incentivized to perform certain actions necessary for the attacker's intent to continue. This activity can be combined with social engineering techniques. At this step, the anti-spam protections of the e-mail servers work. Also here, firewalls can identify the addresses that have access to the system.

4. Exploitation

In this step, the third-step malicious code is executed on the victim's system(s) using the victim's vulnerability. Here and in the next step, antivirus protection works, as well as IDS (Intrusion Detection System) and IDPS (Intrusion Detection and Prevention System) - https://ids-hogzilla.org/, https://www.ossec.net/.

5. Installation

Immediately after the exploit step, the malware or other attack vector is installed on the victim(s) system. This is the most critical point in the attack lifecycle. The threat has entered the system and is taking control of the victim/s.

6. Command and control

In this step, the attacker uses malware or an identity on the target network to take remote control of the device(s). It can also expand their access and establish more entry points for future attacks, increase their privileges and access to resources in the information system.

7. Actions on the goal

At this step, the attacker carries out the intended goals - data theft, destruction, encryption or exfiltration (a low-level attack against DNS servers to gain unauthorized access).

Additionally, some experts extend the chain of cyberattacks to include an eighth step:

Monetization. In this step, the attacker focuses on extracting income from the cyberattack through some form of ransom from the victim or the sale of sensitive information such as personal data or trade secrets.

General Scheme for Data Analysis

Data analysis follows the same pattern as many physical laws are established: collection of experimental data, tabular organization, and establishment of physical laws. Moreover, it is clear that our knowledge of the analyzed process is only approximate. In general, every real-world reasoning system implies different approximations. The analyzed process may turn out to be too complex and do not amenable to analysis using rigorous analytical methods. However, one can gain insight into its behavior under different circumstances by approaching the task from different perspectives, taking into account knowledge of the research domain, experience, intuition, and various heuristics. The movement is from rougher methods to more and more accurate representations of the analyzed process. The general scheme of operation is given in fig. 1. It sets the continuous improvement of the created data model in the presence of new ones. This is how known laws up to now in the field of the fundamental natural disciplines of physics, chemistry and biology were obtained and specified.

This approach requires:

- to consider the problem from different points of view and combine varios approaches;
- not to require high accuracy, to move from the more elementary and rougher models to more complex and more accurate;
- striving for an acceptable result, not an ideal model;
- over time and with the accumulation of new information, the cycle repeats itself - the process of knowledge is endless.

Through the described approach, tasks of acceptable quality are solved. There are flaws in the methodology, but in reality there is no real alternative to it. In the field of natural disciplines, this analysis procedure has been used for many centuries, so there is no reason why it should not be used in other fields as well. The scheme in fig. 1 for data analysis and the models explaining their behavior are also applied to data obtained in cyberattacks.

Cyber security measures encompass three processes: *prevention*, *detection* and *response*.

The use of single-factor and multi-factor authentication of user accounts, encryption of information, firewalls, anti-virus programs, honeypots and others are prevention systems.

Intrusion Detection Systems, Intrusion Detection and Prevention Systems, vulnerability scanners and others are systems for detecting cyber attacks.

The response is determined by the assessed security requirements of the system: upgrade protection, notification of legal authorities, counterattacks and others. In some cases, destroying the compromised system is preferred.

The diagram in Fig. 1 begins with "*Collect experimental data*" consists of distributed storage of experimental data related to cyber attacks and includes:

- Cyber Assets;

- Vulnerabilities;
- Common Vulnerabilities and Exposures (https://www.cve.org/);
- Cyber Threats and Attacks;
- Published Cyber Threats & Attacks;
- Past Data;
- Known Attack Database.

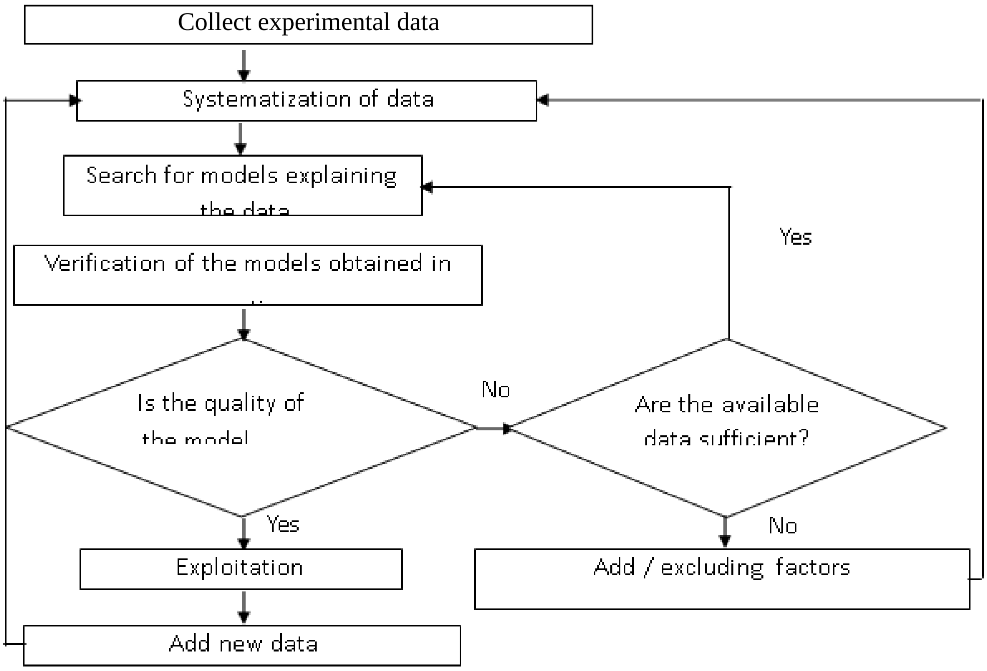
Collect experimental data
Systematization of data
Search for models explaining the data
Verification of the models obtained in
Is the quality of the model
No
Are the available data sufficient?
Yes
Yes
No
Exploitation
Add / excluding factors
Add new data

Fig. 1 General data analysis scheme.

CUBER ATTACK MODELS

They can be known or unknown, so there are two countermeasures: *abuse detection* and *anomaly detection*.

Abuse detection is signature based. It uses patterns of known attacks already being carried out to identify them. A signature is actually a pattern representing a known attack or threat. Abuse detection works by looking for the traces or patterns of known attacks. When a pattern match is found, an event is signaled. An *advantage* is the use of known cyberattacks, their patterns and possible scenarios. A *disadvantage* of abuse detection is that it can only detect attacks that follow known patterns, but is unable to detect future (unknown) attacks that do not have a matching pattern. This technique has a lower false positive rate of anomaly detection, but cannot detect zero-day attacks.

Anomaly Detection is based on the deviation from a known behavior or profile. Stores the baseline of information and communication system or network behavior and data about it such as network traffic, protocols, packet sizes, normal behavior profiles for users, programs, or other resources. An anomaly is a variation or departure from known normal behavior.

A baseline of normal behavior is defined by the system administrator, and a profile is expected or normal behavior obtained by monitoring regular activities, network connections, and hosts, network routers, and users over a period of time. Normal usage patterns can be established using statistical methods when auditing the system for network and other data.

There are three types of anomalies: *in timing*, *number*, and *pattern(s)*.

- *Timing anomalies* occur when activities are performed at unexpected times. To detect these anomalies, a baseline time must be established for all activities of each user and host on the network based on their behavior. It then checks whether the observed activities occur at unusual times compared to the baseline. If deviations occur outside the preset threshold, these are likely anomalies.
- *Count anomalies* occur when an unusually large number of activities are performed in a short period of time by a user and/or host. To detect count anomalies, a baseline must first be established for the count of all system activity performed by users and/or the host. If the observed count is higher than the baseline, then the count is abnormal.
- *Model anomalies* occur when there is an unexpected series of events. Each of these events may not seem anomalous when considered in isolation. However, when viewed together in a sequence and it deviates from what is expected, it becomes a model anomaly.

A model baseline needs to be established. All sequences of events and observed activities are then compared to this baseline. Unlike time and count anomalies, there is no need to consider anomaly models for users and hosts separately.

Examples of model anomalies are:

- A user account installs software on a host at an unusual time: The analyzed pattern is: *Host name* > *Username* > *Time*. The combination of hostname and username is not unusual, but the installation time is, that is, it is an anomaly.
- An unexpected user reads a file containing sensitive host information through a USB port at a specific time: The pattern being analyzed is: *Host name* > *Username* > *Event ID* > *Time*. The *Host name* > *Username* event is anomalous, even though everything else is normal.
- A firewall rule was changed at an unexpected time: The analyzed pattern is: *Host name* > *Rule ID* > *Time*. The *Rule ID* > *Time* event is anomalous.

Abuse and anomaly detection models require processing and analyzing data and extracting dependencies from it. They are the basis for registering cyber attacks. General stages in data mining based cyber-attack detection are:

Data processing:

- the information infrastructure for continuous monitoring and data collection;
- data pre-processing.

Data Analysis:

- search for correlation of events and extraction of functional dependencies;
- data mining.

Visualization and interpretation.

The information infrastructure for continuous monitoring and data collection includes: Firewalls, Antivirus programs, Honeypots, Intrusion Detection System, Intrusion Prevention System, Sniffers (computer software or hardware that can intercept and log traffic passing over a digital network and analyze them), Log Capturers, Security Scanners, Vulnerability Scanners, Malware Detection, Security information and event management.

During data preprocessing Cleansing, Normalization, Collation and others are performed. Main key techniques for processing and mining data for cyber security are [3], [4], [9], [10], [14], [15], [16]:

Predictors:

- Classification;
- Regression;
- Time series analysis;
- Forecast.

Descriptive:

- Clustering;
- Summarizing;
- Associative rules;
- Sequence of patterns.

Common classification techniques include decision tree, K-nearest neighbour classifier, Naive Bayes classifier, support vector machine and such as fuzzy logic, genetic algorithm and neural network.

Clustering is a data mining technique that groups similar items to obtain meaningful groups/clusters of data items in a dataset. Clusters are dominant behavior modes of data objects determined using similarity measures. Clustering provides an effective solution for detecting expected and unexpected behavior patterns and for gaining an understanding of network traffic.

In the process of working, models to counter cyberattacks are trained. A machine learning approach usually consists of two phases: *training* and *testing* [7], [15], [16]. It is performed in the following steps:

- Determination of attributes (features), training and testing set and classes of training data;
- Setting a subset of the attributes needed for classification (dimensionality reduction);
- Creating and training the model using the training data;
- Applying the training model to the testing data.

There are two main types of machine learning: *supervised* and *unsupervised*. In the first case, if the data are of assigned classes (labeled), the task is to find a function or model that explains the data. In the second case, the data is unlabeled and the main task is to find structures, patterns or knowledge in unclassified data.

When detecting abuses in the training phase, each abuse class is trained by using appropriate atributes from the training subset. When testing is performed, new observed data are classified according to whether they belong to some abuse class. If the tested data does not belong to any of the abuse classes, it is classified as normal.

When the observed data belongs to one of the profile classes, it is classified as normal, that is, there is no anomaly. If the observed data does not belong to any of the profile classes, then it is anomalous. It means there is something unexpected. Determining sophisticated attacks requires monitoring anomalies and correlations between them to reveal attack plans.

Applying data processing and analysis as well as extracting dependencies from them in the field of cyber security has some advantages and disadvantages [3], [4], [9]. The following can be mentioned as *advantages*:

- Deep understanding of existing data;
- Identification of security gaps;
- Detection of zero-day attacks (A zero-day attack occurs when hackers exploit recently discovered security flaws before developers have had a chance to fix them. Developers have just learned about the flaw, that is, they are "zero days' to correct it.);
- Detection of sophisticated and masked attack patterns.

Disadvantages include:

- Need for in-depth knowledge in the field of databases, including knowledge of various areas of information and communication technologies such as wired and

wireless hardware, system and application software, protocols, vulnerabilities, threats and others;
- Time and effort to prepare databases for analysis and dependency extraction;
- Constant updating of classified data and techniques for analysis and extraction of dependencies;
- Risk of disclosure of sensitive information in databases;
- Manually inspect the results of the obtained data dependencies.

TYPES OF PENETRATION

Main types of intrusions [14], [15], [17] of cyberattacks are:

- Cyberterrorism - *Stuxnet* is a network worm that infected Windows computers in 2010. It affected Siemens controllers used by Iran to manage uranium enrichment. It can be used for unauthorized data collection (espionage) and sabotage of industrial enterprises, power plants, airports, etc.;
- Insider threats - detecting insider threats is usually a difficult task because these actions often look similar to ordinary user activities. Like intrusion detection systems, insider threat detection systems rely on user profiles to identify legitimate, illegal, and threatening actions.;
- External attacks - to minimize damage from them, as well as from other types of intrusions, penetration testing is performed (ethical intrusion performed by a certified ethical hacker only with the express consent of the institution's management);
- Theft of credit card data, of personal data in order to extract financial benefits;
- Attacks against critical infrastructures – on July 19, 2024, a faulty update by the information security firm CrowdStrike caused a crash of WIDOWS 10 and 11. This led to the blocking of the operation of critical infrastructures such as banks and airports. There are various possibilities for the faulty update, but one of them is a cyber attack on critical infrastructures.

Using data analysis and dependency mining techniques, they can detect financial fraud, telecommunications fraud, computer intrusions, and more. Machine learning is particularly useful for fraud detection. It is necessary to take into account the changes in the number and complexity of the database:

- Detection and prediction of new types of fraud;
- Precise calculation of the probability of fraudulent activity;
- Using supervised (tutored learning) and unsupervised (untutored learning) machine learning algorithms.

With supervised learning, all available records are classified as deceptive or non-deceptive. Tutorless learning training methods create fraud models from unknown records, create their own classification and feature descriptions for fraudulent activities. On this basis, they analyze and detect new types of fraud.

In summary, data dependency mining in computer security applies to:

- Malware detection;
- Intrusion detection;
- Fraud detection;
- Collection of threat information;
- Detection and prediction of insider threats.

Some Multipurpose Cybersecurity Data Analysis Tools

Some key tools for data processing and analysis in the field of security are given bellow.

At *https://www.manageengine.com/log-management* is a *Log360* suite for logging anomalies, abuses, and more.

At *https://desowin.org/USBpcap/* is a *USBPcap* package for capturing packets exchanged over a USB interface for WINDOWS. The content of the same can be viewed and analyzed with the *Wireshark* tool - *https://www.wireshark.org*.

Also *at https://www.cisa.gov/downloading-and-installing-cset* is a *CSET* (Cyber Security Evaluation Tool) package with open source licenses of CISA (Cybersecurity & Infrastructure Securiyt Agensy of USA). CISA also provides a catalog of known exploited vulnerabilities of various operating systems and software products - *https://www.cisa.gov/known-exploited-vulnerabilities-catalog*.

Additionally, *https://www.metasploit.com/download* provides the open source *Metasploit Framework* for logging vulnerabilities and penetration. Other products for work and analysis in the field of information security are indicated at the address (*https://www.rapid7.com/products/insightvm/ and others*).

Address *https://curlie.org/en/Computers/Security* contains over 30 categories of software products and security articles including logging and intrusion prevention systems, firewalls, honeypots for hosts and networks (hijacking cyberattacks), cryptography and others.

Kali Linux (*https://www.kali.org/*) operating system is designed for network analysts, penetration testers, cyber security and analysis experts. It supports more than 600 penetration testing tools like Nmap, Burp Suite, Wireshark, Metasploit Framework, AirCrack-ng, Hydra, WPscan, SQLmap, Nessus and many more. These tools are used for hacking and also for penetration testing. It is a completely free operating system and it is open source.

The Parrot Linux operating system (*https://parrotsec.org/*) is another one that also works in the field of cyber security.

Conclusion

The article presents a basic scheme for processing and analyzing data, machine learning, and on this basis creating data models. Technological concepts of cyber security based on data processing and analysis are summarized and developed. The steps of cyber attacks (also valid for military attacks), use of key techniques of data analysis in cyber defense, application of data processing and analysis in information security and tools used are indicated.

The material provided in the article can be considered not only as an application of data processing and analysis with subsequent extraction of dependencies in the field of cyber security and creation of corresponding behavior models, but also as a basic guide for work in this area, but also development guidelines.

Concrete applications of data analysis, dependency extraction, machine learning, creation of relevant models in the field of cyber defense is a future field of research in theoretical and practical aspects (software).

References // Литература

1. ENISA , "Home", Available at: https://www.enisa.europa.eu (last view: 15-07-2024)
2. Law on cyber security, pub. SN. No. 94 of November 13, 2018, Available at: https://lex.bg/bg/laws/ldoc/2137188253 (last view: 15-07-2024)
3. Farhad Foroughi, Peter Luksch, "Data Science Methodology for Cybersecurity Projects", Institute Of Computer Science University Of Rostock, Rostock, Germany, Available at: https://www.researchgate.net/ (last view: 15-07-2024), 2018.

4. Leslie F. Sikos, Kim-Kwang Raymond Choo, “Data Science in Cybersecurity and Cyberthreat Intelligence”, Springer Nature Switzerland AG 2020, ISBN 978-3-030-38788-4 (eBook), DOI: https://doi.org/10.1007/978-3-030-38788-4
5. MITRE ATT&CK, “ATT&CK”, Available at: https://attack.mitre.org/ (last view: 15-07-2024)
6. NCIA, “NATO Cyber Security Centre” Available at: https://www.ncia.nato.int/what-we-do/cyber-security.html (last view: 15-07-2024)
7. KDnuggets, “Data Science, Machine Learning, AI & Analytics”, Available at: https://www.kdnuggets.com/ (last view: 15-07-2024)
8. NEC, “How data mining techniques can be used in cyber security solutions”, Available at: https://www.nec.co.nz/market-leadership/publications-media/how-data-mining-techniques-can-be-used-in-cyber-security-solutions/ (last view: 15-07-2024)
9. Apriorit, “Using Data Mining Techniques in Cybersecurity Solutions” Available at: https://www.apriorit.com/dev-blog/527-data-mining-cyber-security#q (last view: 15-07-2024)
10. Orient Software Development Corp., “Learning the Basics of Big Data Analytics for Cybersecurity” Available at: https://www.orientsoftware.com/blog/data-analytics-cybersecurity/ (last view: 15-07-2024)
11. Faisal A. Garba, “The Anatomy of a Cyber Attack: Dissecting the Cyber Kill Chain”, Available at: https://www.researchgate.net/publication/330658097 (last view: 15-07-2024).
12. Trupti Zaware, “Cybersecurity Automation Using Cyber Kill Chain”, International Research Journal of Modernization in Engineering Technology and Sciencee-ISSN: 2582-5208, Volume:04/Issue:09/, DOI: https://www.doi.org/10.56726/IRJMETS29648 www.irjmets.com, September-2022.
13. Lockheed Martin Corporation, “Cyber Kill Chain” Available at: https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html (last view: 15-07-2024)
14. Vivek Ramachandran, Cameron Buchanan, “Kali Linux Wireless Penetration”, Published by Packt Publishing Ltd Birmingham, ISBN 978-1-78328-041-4, www.packtpub.com, March 2015.
15. Sayan Mukhopadhyay, “Advanced Data Analytics Using Python”, Kolkata, West Bengal, India, ISBN-13 (electronic): 978-1-4842-3450-1, DOI: https://doi.org/10.1007/978-1-4842-3450-1 2018.
16. Shterev Yordan, “Data analysis”, Publ. Faber, Veliko Turnovo, ISBN 978-954-400-321-0, 2010.
17. Thomas J. Mowbray, “Cybersecurity: Managing Systems, Conducting Testing, and Investigating Intrusions”, Published by John Wiley & Sons, ISBN: 978-1-118-84965-1 (ebk), 2014.

Received: 25-07-2024 Accepted: 25-09-2024 Published: 20-12-2024

Cite as:
Shterev, Y. (2024). “Data Mining in Cybersecurity”, Science Series “Innovative STEM Education”, volume 06, ISSN: 2683-1333, pp. 99-107, 2024. DOI: https://doi.org/10.55630/STEM.2024.0610