Chapter 2
Cybersecurity Data Science: Concepts, Algorithms, and Applications

Wei Lu

2.1 Introduction

Cybersecurity is an ability to protect or defend the use of cyberspace from cyberattacks. The four most common goals provided by cybersecurity technologies for any users/entities in cyberspace are confidentiality, authentication, message integrity, and access and availability [1]. Confidentiality ensures that the information is inaccessible to unauthorized people, and it is commonly enforced through encryption, IDs and passwords, two-factor authentication, and additional defensive strategies in which only sender and the intended receiver should understand message contents. In the context of confidentiality, the sender encrypts the message and the receiver decrypts the message. A typical example of confidentiality is student grade information that is an asset whose confidentiality is important to students, and only be available to students, their parents, and employees that require the information to do their job. Message integrity is to safeguard information and systems from being modified by unauthorized people, thereby ensuring that the protected data is accurate and trustworthy. Considering, for example, a hospital patient's allergy information stored in a database and trusted by the doctor, the integrity is broken when an employee (e.g., a nurse) who is authorized to view and update this information deliberately falsified the data to cause harm to the hospital. Then, the database needs to be restored to a trusted basis quickly, and trace the error back to the person responsible. In the context of message integrity, sender and receiver want to ensure messages are not altered in transit, or afterwards

W. Lu (✉)
Department of Computer Science, Keene State College, The University System of New Hampshire, Keene, NH, USA
e-mail: wlu@usnh.edu

I. Woungang, S. K. Dhurandher, *4th International Conference on Wireless, Intelligent and Distributed Environment for Communication*, Lecture Notes on Data Engineering and Communications Technologies 94, https://doi.org/10.1007/978-3-030-89776-5_2

without detection. Availability is to ensure that authorized people have access to the information when needed; this includes rigorously maintaining all systems, keeping them current with upgrades, using backups to safeguard against disruptions or data loss, to name a few. For example, a system providing authentication services for critical systems, applications, and devices. An interruption of service results in the inability for customers to access computing resources and staff to access the resources they need to perform critical tasks. The loss of the service translates into a large financial loss in lost employee productivity and potential customer loss. In the context of availability, services must be accessible and available to users 24 h 7 days.

The early cybersecurity mechanisms dedicated exclusively to prevention are highly insufficient mainly because current Internet activities have shown a dramatic increase in the number of computer and network attacks [2]. It was estimated that cybercrime is the second most-strongly connected global risk, and the cost of its damage would be equivalent to the GDP of the world's third largest economy in 2021 [3]. Under current circumstances, password-based authentication and access control mechanisms, which represent the cornerstone of traditional protection systems, can easily be circumvented using widely available exploits [4]. As a result, the concept of cybersecurity data science is proposed to supplement traditional prevention-based security mechanisms.

In the rest of this short chapter, we introduce the basic concepts of cybersecurity data science in Sect. 2.2 including what is cybersecurity, information security and network security, and what are their main differences, what is data science, and what is cybersecurity data science. In Sect. 2.3, we review an unsupervised machine learning algorithm [5], and then in Sect. 2.4, we present how to use the proposed unsupervised machine learning algorithm for detecting zero-day attacks, a typical case study on cybersecurity data science. Section 2.5 makes some concluding remarks and discusses future work.

2.2 Concepts of Cybersecurity Data Science

Cybersecurity, information security and network security are three terms that are often used interchangeably, even among some of those in the security field. However, they are not exactly the same. Each of them addresses different kinds of security. Understanding each term, what it means, and the difference among the three, i.e., what are they, how are they different, and why are these terms so often confused, is basic and essential for any organization that is investing in a proper security framework.

We have known that cybersecurity is an ability to protect or defend the use of cyberspace from cyberattacks, while computer security is a set of measures and controls that ensure confidentiality, integrity, and availability of the information processed and stored by a computer [6]. This term has been replaced by the term cybersecurity. On the other hand, information security is the protection of

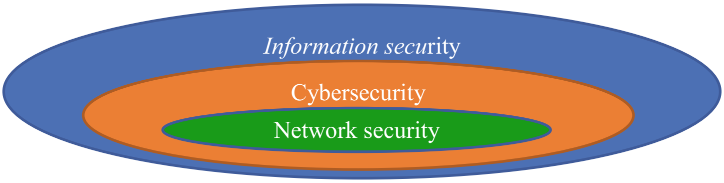

Fig. 2.1 Information security vs. cybersecurity vs. network security

information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction in order to provide confidentiality, integrity, and availability, and network security represents the process of taking physical and software preventative measures to protect the underlying networking infrastructure from unauthorized access, misuse, malfunction, modification, destruction, or improper disclosure, thereby creating a secure platform for computers, users, and programs to perform their permitted critical functions within a secure environment.

As illustrated in Fig. 2.1, information security is, broadly, the practice of securing your data, no matter its form, and cybersecurity is a subset of information security that deals with protecting an organization's internet-connected systems from potential cyberattacks; moreover, network security is a subset of cybersecurity that is focused on protecting an organization's IT infrastructure from online threats. Although the terms are often used in conjunction with one another, cybersecurity is considered to be the broader discipline as long as data originates in a digital form.

2.2.1 What Is Data Science?

Generally speaking, data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from many structural and unstructured data. As illustrated in Fig. 2.2, the Data Science Venn Diagram includes six components, namely Data Hacking Skills, Math and Statistics Knowledge, Machine Learning, Traditional Research, Danger Zone, and Substantive Expertise (domain knowledge) [7].

- *Data Hacking Skills*: Data hacker/engineer with skills to acquire and clean data, thinking algorithmically, being able to manipulate data using command line, understanding vectorized operations, to name a few.
- *Math and Statistics Knowledge*: Extract insight from the data, at least needs to know what is least squares regression to be a competent data scientist.
- *Substantive Expertise (Domain Knowledge)*: Substantive expertise is essential and is built upon domain knowledge, because data science is about discovery and building knowledge, which at least requires some motivating questions about the world and hypotheses that can be brought to data and tested with statistical methods.

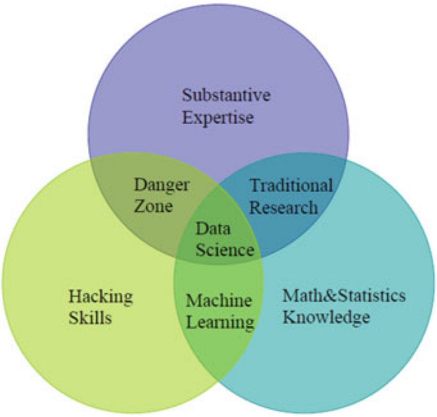

Fig. 2.2 The data science Venn diagram [7]

- *Machine Learning*: Data plus and math/statistics get us to machine learning.
- *Traditional Research*: Domain knowledge and math/statistics get you traditional research.
- *Danger Zone*: Data plus domain knowledge is sparsely populated and is the most problematic of the diagram, could produce a lot of false interpretation without learning some math and statistics along the way.

As a result, data science is an interdisciplinary field that combines Data Hacking Skills, Math and Statistics Knowledge, and Substantive Expertise (Domain Knowledge).

2.2.2 What Is Cybersecurity Data Science?

Cybersecurity data science is an interdisciplinary field that combines data hacking skills, math and statistics knowledge, and substantive expertise in cybersecurity including in particular one or more of the following subjects:

- SIEM development
- Insider threat detection
- Computer and network forensics
- Security metrics; governance, risk, and compliance
- Risk modeling; fraud and loss analytics
- Advanced threat mitigation; malware analysis

2.3 Algorithms for Cybersecurity Data Science

The EM algorithm is widely used to estimate the parameters of Gaussian Mixture Model (GMM) [8]. GMM is based on an assumption that the data to be clustered are drawn from one of several Gaussian distributions. It is suggested that Gaussian mixture distributions can approximate any distribution up to an arbitrary accuracy, as long as a sufficient number of components are used. Consequently, the entire data collection is seen as a mixture of several Gaussian distributions, and their corresponding probability density functions can be expressed as a weighted finite sum of Gaussian components with different parameters and mixing proportions.

The conditional probability in EM describes the likelihood that data points approximate a specified Gaussian component. The greater the value of conditional probability for a data point belonging to a specified Gaussian component, the more accurate the approximation is. As a result, data points are assigned to the corresponding Gaussian components according to their conditional probabilities. However, in some cases, there exist some data points whose conditional probability of belonging to any component of a GMM is very low or close to zero. These data are naturally seen as the outliers or noisy data. All the outlier data will be deleted or considered as unknown (aka zero-day) attacks during anomaly detection, and their attacking probability is set to 1.0. Table 2.1 illustrates a detailed EM-based clustering algorithm in which C_m stands for the clustering results.

In order to apply the EM-based clustering technique for detecting zero-day attacks, we make two basic assumptions: (1) the input data points are composed of two clusters, namely anomalous cluster and normal cluster; (2) the size of the anomalous cluster is always smaller than the size of the normal cluster. Consequently, we can easily label the anomalous cluster according to the size of each cluster. The attack probability for each data point is equal to the conditional probability of corresponding data point belonging to the anomalous cluster, which is defined as follows:

$$p = p_{r-1}\left(C_{\text{anomalous}} | x_n\right)$$

where x_n is the data point; $C_{\text{anomalous}}$ is the anomalous cluster; $p_{r-1}(C_{\text{anomalous}} | x_n)$ is the conditional probability of x_n belonging to anomalous cluster $C_{\text{anomalous}}$.

Table 2.1 EM-based clustering algorithm

```
Function EMCA (data) returns
  clusters $C_m$ and posterior probability $p_r(i | x_n)$
    $C_m = \phi$, 1 £ $m$ £ $k$, $k$ is the number of clusters
    Call EM (data);
    For 1 £ $m$ £ $k$, 1 £ $n$ £ $N$
      If ( $p_{r-1}(m | x_n) = \max(p_{r-1}(m | x_n))$ )
      Then assign $x_n$ to $\mathrm{C_m}$
    Return $C_m$, $m = 1,2...,k$
```

2.4 Applications for Cybersecurity Data Science

In order to apply the proposed EM-based clustering algorithm for detecting zero-day attacks in cybersecurity data science as a case study, we develop an effective and efficient clustering framework for online unsupervised anomaly detection of unknown intrusions. The proposed detection framework consists of a feature extraction technique based on the number of network flows and packets and an EM-based unsupervised machine learning algorithm. Figure 2.3 depicts the general architecture of our framework, which consists of three main stages as follows:

1. Feature analysis: During this phase, a number of flows and packets are generated from standard packet information, allowing extraction of salient and useful domain knowledge, as well as significant reduction of the dimensionality in the feature space.
2. Clustering: The features computed during the previous step are clustered using the EM-based clustering algorithm.
3. Cluster analysis and intrusion decision: Analyzing the clustering outcomes using heuristics leads to a final classification of corresponding data as either unknown novel attacks or nonintrusive data.

The major goal of feature analysis is to select and extract robust network features that have the potential to discriminate anomalous behaviors from normal network activities. Since most current network intrusion detection systems use network flow data (e.g., netflow, sflow, ipfix) as their information sources, we focus on features in terms of flows.

FlowCount A flow consists of a group of packets going from a specific source to a specific destination over a time period. There are various flow definitions so far, such as netflow, sflow, ipfix, to name a few. Basically, one network flow should at least include source IP/port, destination IP/port, protocol, number of bytes, and number of packets. Flows are often considered as sessions between users and services. Since attacking behaviors are usually different from normal user activities, they may be detected by observing flow characteristics.

PacketCount The average number of packets in a flow over a time interval. Most attacks happen with an increased packet count. For example, Distributed Denial of

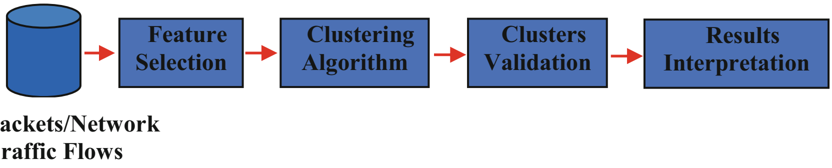

Fig. 2.3 A typical cybersecurity data science application framework

Table 2.2 List of features

Features	Description
f_1	Number of TCP flows per minute
f_2	Number of UDP flows per minute
f_3	Number of ICMP flows per minute
f_4	Average number of TCP packets per flow over 1 m
f_5	Average number of UDP packets per flow over 1 m
f_6	Average number of ICMP packets per flow over 1 m
f_7	Average number of bytes per TCP flow over 1 m
f_8	Average number of bytes per UDP flow over 1 m
f_9	Average number of bytes per ICMP flow over 1 m

Service (DDoS) attacks often generate a large number of packets in a short time in order to consume the available resources quickly.

ByteCount The average number of bytes in a flow over a time interval. Through this metric, we can identify whether the network traffic consists of large size packets or not. Some previous Denial of Service (DoS) attacks use maximum packet size to consume the computation resources or to congest data paths, such as the well-known ping of death (pod) attack.

Based on the above three metrics, we define a set of features to describe entire traffic behavior on networks. Let F denote the feature space of network flows, a nine-dimensional feature vector $f\in$ can be represented as $\{\boldsymbol{f_1},\boldsymbol{f_2}, \ldots, \boldsymbol{f_9}\}$, where the meaning of each feature is explained in Table 2.2.

We evaluate the application framework with the full 1999 DARPA intrusion detection dataset. In particular, we conduct a comprehensive analysis for network traffic provided by the dataset and identify the intrusions based on each specific day. Since most current existing network intrusion detection systems use network flow data (e.g., netflow, sflow, ipfix) as their information sources, we covert all the raw TCPDUMP packet data into flow-based traffic data by using the public network traffic analysis tools (e.g., editcap [9], tshark [10]), similarly to the 1999 KDDCUP dataset [11] in which the 1998 DAPRA intrusion detection dataset [12] has been converted into connection-based dataset. Although the 1998 and 1999 DARPA dataset were criticized in [13, 14] due to the methodology for simulating the actual network environment, they are the widely used and acceptable benchmark for the current intrusion detection research.

During the evaluation, the results are summarized and analyzed in three different categories, namely how many attack instances are detected by each feature, how many attack types are detected by each feature, and how many attack instances are detected for each attack type. We do not use the traditional receiver operating characteristic (ROC) curve to evaluate our approach and analyze the tradeoff between the false-positive rates and detection rates because ROC curves are often misleading and incomplete. Compared to most, if not all, other evaluations with the 1999 DARPA dataset, our evaluation covers all types of attacks and all days'

Table 2.3 Performance of all nine features over 9 days evaluation for EM clustering

Features	Average DR (%)	Average FPR (%)	Ratio of Avg. DR to Avg. FPR
F1	39.83	81.84	0.487
F2	52.22	84.04	0.621
F3	32.25	84.14	0.383
F4	12.0	89.03	0.135
F5	51.8	85.74	0.604
F6	32.25	84.17	0.383
F7	3.2	82.92	0.0386
F8	49.26	84.19	0.585
F9	32.25	84.14	0.383

network traffic, and thus, we consider our evaluation as a comprehensive analysis for network traffic in the 1999 DARPA dataset.

The scoring coefficient is set up according to the detection rate (DR) and the false-positive rate (FPR) for the developed system over a long history. The higher the DR, the better the performance of the system; the lower the FPR, the better the system performance. Therefore, the ratio of DR to FPR is used to measure the performance of our proposed framework. We evaluate our system with the nine features and 9 days DARPA testing data on weeks 4 and 5. The evaluation results are summarized and analyzed in three different categories described in above. Table 2.3 illustrates the average value of DR, FPR, and the ratio of DR to FPR for each feature over those 9 days. From Table 2.3, we can see that the top three features that are extremely useful for detecting zero-day attacks are number of UDP flows per minute, average number of UDP packets per flow over 1 min, and average number of bytes per UDP flow over 1 min.

2.5 Conclusions and Future Work

In this short chapter, we discuss the essential concepts of cybersecurity data science and its typical algorithms and applications in detecting novel network attacks. The concept of cybersecurity data science is not new and can be originally dated back to the 1980s in the seminal report of Denning [15]. Denning assumed that security violations could be detected by inspecting abnormal system usage patterns from the audit data. Deviations from normal behavior patterns are flagged systematically as intrusions. The implementations of early anomaly detection techniques were based on self-learning. Knowledge about normal behaviors of subjects was automatically formed through training. Thus, according to whether the learning process is supervised or unsupervised, the anomaly detection schemes are naturally classified into two categories: unsupervised and supervised.

Supervised anomaly detection schemes depend on labeled training datasets, making the intrusion detection process error-prone, costly and time consuming.

Any mistake in labeling the training data may lead to decreased performance of the detector. Unsupervised anomaly detection schemes allow training based on unlabeled datasets, facilitating online learning and improving detection accuracy. By facilitating online learning, unsupervised approaches provide higher potential to find novel attacks, which are not always included in the training data [16–18]. By removing the need to label the dataset, unsupervised approaches carry greater potential for detection accuracy, and they, however, also carry greater computing overheads. In the future, we will have supervised learning, in particular, a traditional perceptron learning algorithm, and unsupervised learning approaches integrated into the proposed cybersecurity data science application framework in order to bridge the knowledge gap between fields so that legal practitioners, law enforcement agencies, policy makers, and economists can make wise choices when dealing with societal problems related to cybersecurity as well as its broad economic impact.

Acknowledgments This research was supported in part by funding from a Keene State College Faculty Development Grant.

References

1. Stallings, W. (2019). *Cryptography and network security: Principles and practice* (8th ed.). Pearson. ISBN-13: 978-0135764183.
2. Ghorbani, A. A., Lu, W., & Tavallaee, M. (2010). Network attacks. In *Network intrusion detection and prevention. Advances in information security* (Vol. 47). Springer.
3. The Global Risks Report 2020. (2021). https://www.weforum.org/reports/the-global-risks-report-2020
4. Lu, W., & Traore, I. (2005). An unsupervised approach for detecting DDOS attacks based on traffic-based metrics. In *2005 IEEE pacific rim conference on communications, computers and signal processing* (pp. 462–465).
5. Lu, W., & Traore, I. (2005). Determining the optimal number of clusters using a new evolutionary algorithm. In *17th IEEE international conference on tools with artificial intelligence (ICTAI'05)*.
6. National Institute of Standards and Technology. (2015). https://csrc.nist.gov/glossary/
7. The Data Science Venn Diagram. (2013). http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram
8. Lu, W., & Traore, I. (2005). A new evolutionary algorithm for determining the optimal number of clusters. In *International conference on computational intelligence for modelling, control and automation and international conference on intelligent agents, web technologies and internet commerce* (Vol. 2005, pp. 648–653).
9. Editcap. https://www.wireshark.org/docs/wsug_html_chunked/AppToolseditcap.html
10. Tshark. https://www.wireshark.org/docs/man-pages/tshark.html
11. KDDCUP. http://archive.ics.uci.edu/ml/datasets/KDD+Cup+1999+Data
12. *The 1998 DARPA intrusion detection dataset*. https://www.ll.mit.edu/r-d/datasets/1998-darpa-intrusion-detection-evaluation-dataset#:~:text=Intrusion%20detection%20systems%20were%20delivered,real%20time%20during%20normal%20activities
13. Mahoney, M. V., & Chan, P. K. (2003). An analysis of the 1999 DARPA/Lincoln Laboratory evaluation data for network anomaly detection. In *Proceedings of the 6th international symposium on recent advances in intrusion detection* (pp. 220–237).

14. McHugh, J. (2000). Testing intrusion detection systems: A critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory. *ACM Transactions on Information and System Security, 3*(4), 262–294.
15. Denning, D. E. (1987). An intrusion detection model. *IEEE Transactions on Software Engineering, SE-13*(2), 222–232.
16. Lu, W., & Xue, L. (2014). A heuristic-based co-clustering algorithm for the internet traffic classification. In *2104 28th international conference on advanced information networking and applications workshops* (pp. 49–54). https://doi.org/10.1109/WAINA.2014.16
17. Garant, D., & Lu, W. (2013). Mining botnet behaviors on the large-scale web application community. In *2013 27th international conference on advanced information networking and applications workshops* (pp. 185–190). https://doi.org/10.1109/WAINA.2013.235
18. Lu, W., & Traore, I. (2008). Unsupervised anomaly detection using an evolutionary extension of k-means algorithm. *International Journal on Information and Computer Security, 2*(2), 107–139.