Chapter 14
Big Data Analytics for Intrusion Detection: An Overview

Luis Filipe Dias
https://orcid.org/0000-0003-2842-6655
Instituto Universitário Militar, Portugal

Miguel Correia
Universidade de Lisboa, Portugal

ABSTRACT

Intrusion detection has become a problem of big data, with a semantic gap between vast security data sources and real knowledge about threats. The use of machine learning (ML) algorithms on big data has already been successfully applied in other domains. Hence, this approach is promising for dealing with cyber security's big data problem. Rather than relying on human analysts to create signatures or classify huge volumes of data, ML can be used. ML allows the implementation of advanced algorithms to extract information from data using behavioral analysis or to find hidden correlations. However, the adversarial setting and the dynamism of the cyber threat landscape stand as difficult challenges when applying ML. The next generation security information and event management (SIEM) systems should provide security monitoring with the means for automation, orchestration and real-time contextual threat awareness. However, recent research shows that further work is needed to fulfill these requirements. This chapter presents a survey on recent work on big data analytics for intrusion detection.

INTRODUCTION

Over the past two decades *network intrusion detection systems* (NIDSs) have been intensively investigated in academia and deployed by industry (Debar et al., 1999). More recently, intrusion detection has become a big data problem because of the growing volume and complexity of data necessary to unveil increasingly sophisticated cyberattacks. The *security information and event management* (SIEM)

DOI: 10.4018/978-1-5225-9611-0.ch014

Copyright © 2020, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

systems adopted during the last decade show limitations when processing with *big data*, even more in relation to extracting the information it can provide. Therefore, new techniques to handle high volumes of security-relevant data, along with *machine learning* (ML) approaches, are receiving much attention from researchers. This chapter presents an overview of the state-of-the-art regarding this subject.

The Cloud Security Alliance (CSA) suggested that *intrusion detection systems* (IDSs) have been going through three stages of evolution corresponding to three types of security tools (Cárdenas et al., 2013):

- IDS: able to detect well-known attacks efficiently using signatures (misuse detection) and to unveil unknown attacks at the cost of high false alarm rates (anomaly detection);
- SIEM: collect and manage security-relevant data from different devices in a network (e.g., firewalls, IDSs, and authentication servers), providing increased network security visibility by aggregating and filtering alarms, while providing actionable information to security analysts;
- 2nd generation SIEM: the next generation, that should be able to handle and take the best from big data, reducing the time for correlating, consolidating, and contextualizing even more diverse and unstructured security data (e.g., global threat intelligence, blogs, and forums); they should be able to provide long-term storage for correlating historical data as well as for forensic purposes.

The European Agency for Network and Information Security (ENISA) stated that the next generation SIEMs are the most promising domains of application for big data (ENISA, 2015). According to recent surveys from the SANS Institute, most organizations are just starting to evolve from traditional SIEMs to more advanced forms of security analytics and big data processing (Shackleford, 2015, 2016). In fact, the industry developments in the area led Gartner to start publishing market guides for user and entity behavior analytics (UEBA) (Litan, 2015). While recent UEBA technologies target specific security use cases (e.g., insider threats), typical SIEM technologies provide comprehensive rosters of all security events which are also important for compliance requirements. Recent guides from Gartner (Bussa et al., 2016; Kavanagh et al., 2018) state that "Vendors with more mature SIEM technologies are moving swiftly to incorporate big data technology and analytics to better support detection and response", revealing the tendency of moving towards 2nd generation SIEMs.

The focus of this survey is on state-of-the-art techniques that can contribute for such next generation SIEMs, and on the challenges of big data and ML applied to cybersecurity analytics. There are a few related surveys available in the literature, none with this focus. Buczak & Guven (2016) analyze papers that use different ML techniques in the cybersecurity domain; although interesting, most of the experiments of those studies were done with datasets that date back to 1999 and that do not represent the actual cybersecurity landscape. Bhuyan et al. (2014) provide a comprehensive overview of network anomaly detection methods. Zuech et al. (2015) review works considering the problem of big heterogeneous data associated with intrusion detection. While this last work touches topics similar to this chapter, it lacks a comprehensive study on recent techniques that tackle the problem of extracting useful information from such volumes of data.

When selecting the papers to investigate, we prioritized recent work – less than 5 years old – and those that report experiments with real-world data. The rationale is that older approaches and approaches evaluated with synthetic datasets (e.g., KDD99) may be inadequate to detect real/recent attacks. Furthermore, we restricted the focus to papers published in major conferences/journals, with few exceptions.

23 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/big-data-analytics-for-intrusion-detection/235047

Related Content

Reasoning Temporally Attributed Spatial Entity Knowledge Towards Qualitative Inference of Geographic Process

Jayanthi Ganapathyand Uma V. (2019). *International Journal of Intelligent Information Technologies (pp. 32-53).*

www.irma-international.org/article/reasoning-temporally-attributed-spatial-entity-knowledge-towards-qualitative-inference-of-geographic-process/225068

Mehar Approach for Solving Shortest Path Problems With Interval-Valued Triangular Fuzzy Arc Weights

Tanveen Kaur Bhatia, Amit Kumar, M. K. Sharmaand S. S. Appadoo (2022). *International Journal of Fuzzy System Applications (pp. 1-17).*

www.irma-international.org/article/mehar-approach-for-solving-shortest-path-problems-with-interval-valued-triangular-fuzzy-arc-weights/313428

Exploring the Profound Impact of AI on Higher Education and Students: Shaping Tomorrow's Workforce

Amy Emanueland David H. Stone (2024). *Academic Integrity in the Age of Artificial Intelligence (pp. 112-138).*

www.irma-international.org/chapter/exploring-the-profound-impact-of-ai-on-higher-education-and-students/339222

Application of Machine Learning for Optimization

Paramita Deyand Kingshuk Chatterjee (2023). *Handbook of Research on AI and Machine Learning Applications in Customer Support and Analytics (pp. 113-127).*

www.irma-international.org/chapter/application-of-machine-learning-for-optimization/323117

Facial Expression Recognition for HCI Applications

Fadi Dornaikaand Bogdan Raducanu (2009). *Encyclopedia of Artificial Intelligence (pp. 625-631).*

www.irma-international.org/chapter/facial-expression-recognition-hci-applications/10312