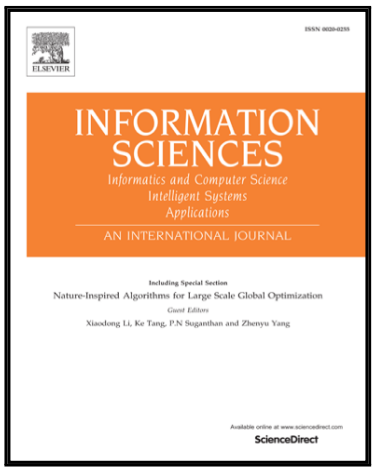

Abstracting Massive Data for Lightweight Intrusion Detection in Computer Networks

Wei Wang, Jiqiang Liu, Georgios Pitsilis, Xiangliang Zhang

PII: S0020-0255(16)31238-5
DOI: 10.1016/j.ins.2016.10.023
Reference: INS 12582

To appear in: *Information Sciences*

Received date: 14 March 2016
Revised date: 3 October 2016
Accepted date: 9 October 2016

Please cite this article as: Wei Wang, Jiqiang Liu, Georgios Pitsilis, Xiangliang Zhang, Abstracting Massive Data for Lightweight Intrusion Detection in Computer Networks, *Information Sciences* (2016), doi: 10.1016/j.ins.2016.10.023

Abstracting Massive Data for Lightweight Intrusion Detection in Computer Networks

Wei Wang [a], Jiqiang Liu [a], Georgios Pitsilis [b], Xiangliang Zhang [c]

[a] *School of Computer and Information Technology, Beijing Jiaotong University*
No.3 Shangyuancun, 100044 Beijing, China

[b] *Computer Science Research, Athens, Greece*

[c] *Division of Computer, Electrical and Mathematical Sciences & Engineering*
King Abdullah University of Science and Technology (KAUST), Saudi Arabia

Abstract

Anomaly intrusion detection in big data environments calls for lightweight models that are able to achieve real-time performance during detection. Abstracting audit data provides a solution to improve the efficiency of data processing in intrusion detection. Data abstraction refers to abstract or extract the most relevant information from the massive dataset. In this work, we propose three strategies of data abstraction, namely, exemplar extraction, attribute selection and attribute abstraction. We first propose an effective method called exemplar extraction to extract representative subsets from the original massive data prior to building the detection models. Two clustering algorithms, Affinity Propagation (AP) and traditional k-means, are employed to find the exemplars from the audit data. K-Nearest Neighbor (k-NN), Principal Component Analysis (PCA) and one-class Support Vector Machine (SVM) are used for the detection. We then employ another two strategies, attribute selection and attribute extraction, to abstract audit data for anomaly intrusion detection. Two http streams collected from a real computing environment as well as the KDD'99 benchmark data set are used to validate these three strategies of data abstraction. The comprehensive experimental results show that while all the three strategies improve the detection efficiency, the AP-based exemplar extraction achieves the best performance of data abstraction.

Key words: Data reduction, intrusion detection, anomaly detection, computer security

1 Introduction

The importance of computer network security is growing with the pervasive involvement of computers in people's daily lives and in business processes within most organizations. As an important technique in the defense-in-depth network security framework, intrusion detection has become a widely studied topic in computer networks in recent years.

In general, the techniques for intrusion detection can be categorized as signature-based detection and anomaly detection. Signature-based detection (e.g., Snort [31]) relies on a database of signatures from known malicious threats. Anomaly detection, on the other hand, defines a profile of a subject's normal activities and attempts to identify any unacceptable deviation as a potential attack. Typically, machine learning techniques are used to build normal profiles of a subject. Any observable behavior of a system, such as a network's traffic [13,19], a computer host's operating system [11,36] or a mobile application [2,39], can be used as the subject information.

Anomaly detection has a potential to detect unforeseen attacks. As new attacks appear very frequently and signature-based detection methods may be overwhelmed by an abundance of polymorphic attacks, using anomaly detection sensors to discover zero-day attacks has become a necessity rather than an option [8]. We are entering the era of "big data" [23]. The increasing volume of information generated by enterprises, the rise of social media and the Internet are fueling an exponential growth of data. Anomaly intrusion detection techniques

Email addresses: wangwei1@bjtu.edu.cn, http://infosec.bjtu.edu.cn/wangwei (Wei Wang), jqliu@bjtu.edu.cn, http://infosec.bjtu.edu.cn/jqliu (Jiqiang Liu), georgios.pitsilis@gmail.com (Georgios Pitsilis), xiangliang.zhang@kaust.edu.sa, http://mine.kaust.edu.sa (Xiangliang Zhang).