Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 61 (2015) 46 – 51

Procedia Computer Science

Complex Adaptive Systems, Publication 5
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2015-San Jose, CA

Using Data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS)

Solane Duque[a]*, Dr.Mohd. Nizam bin Omar[b]

[a]*College of Arts and Sciences,University Utara Malaysia*
[b]*College of Arts and Sciences,University Utara Malaysia*

Abstract

A common problem shared by current IDS is the high false positives and low detection rate. An unsupervised machine learning using k-means was used to propose a model for Intrusion Detection System (IDS) with higher efficiency rate and low false positives and false negatives. The NSL-KD data set was used which consisted of 25,192 entries with 22 different types of data. Results of the study using 11, 22, 44, 66 and 88 clusters, showed an efficiency rate of 70.75%, 81.61%, 65.40%, 61.30% and 55.43% respectively; false positive rates of 0.74%, 4.03%, 15.55%, 21.47% and 31.91% respectively; and false negative rates of 99.82%, 98.14%, 97.76%, 96.32% and 95.70%, respectively. Interestingly, the best results were generated when the number of clusters matches the number of data types in the data set. In the light of the findings, it is recommended that other data mining techniques be explored; a study using k-means data mining algorithm followed by signature-based approach is proposed in order to lessen the false negative rate; and a system for automatically identifying the number of clusters may be developed.

Peer-review under responsibility of scientific committee of Missouri University of Science and Technology

Keywords: data mining; clustering; machine learning; unsupervised learning; k-means

1. Introduction

The latest developments in computer systems and the internet have revolutionized the way people think and do things. A process like sending traditional mail that normally takes hours or even days can now be completed in a click of a mouse or a touch of a finger through electronic mail or e-mail. People communicate with each other from

* Corresponding author. Tel.: +968-9609-7575;
E-mail address: solane@unizwa.edu.om

Peer-review under responsibility of scientific committee of Missouri University of Science and Technology
doi:10.1016/j.procs.2015.09.145

different places through integrated relay chat, or video conferencing as a much convenient mode of communication.

However, along with the many advances in computer systems and IT infrastructures are the risks associated with the use of these technologies. Over the last two decades, computer threats and cybercrimes have proliferated at the disadvantage of the general public, and newer threats are introduced each day that compromise the integrity, validity and confidentiality of data. Companies, nations, and individual persons can be victims of malicious activities in the internet. As a consequence of cybercrimes, millions of dollars have been spent on mitigation strategies.

People who exploit the vulnerabilities of the information systems are usually adept at using sophisticated programming techniques and take advantage of the interconnectivity of the systems so much so that they do not even need local access to the network because they can launch the attacks remotely.

Malicious activities in the internet are also known as intrusion. An intrusion is defined as any activity that violates security policy of the network [1]. Intrusion detection system (IDS) is software and hardware deployed to carry out the process of detecting unauthorized use of, or attack upon, a computer or a telecommunications network – which is supposed to bridge the gaps in firewall and anti-viruses. An IDS provides monitoring and analysis of user and system activity, can audit system configuration and vulnerabilities, assess the integrity of critical system and data files, provide statistical analysis of activity patterns based on the matching with known attacks, analyze abnormal activity, and operate system audit [2]. One advantage of the IDS is its ability to document the intrusion or threat to an organization, thereby providing bases for informing the public regarding the latest attack patterns through system logs.

The types of computer attacks detected by IDS are categorized into three, namely: (i) scanning attacks, (ii) denial of service (DOS) attacks, and (iii) penetration attacks [3]. Each of these three categories of computer attacks has distinct signatures and behaviours - to which IDS is designed to analyze, detect and triggers an alarm when encountered. Once an alarm is set, network administrators will have to analyze the logs to decide whether the suspected activity is indeed anomalous.

In most IDS however, there is a high instances of false positives and false negatives which can be cumbersome to deal with for the network administrators. A false positive is an instance where an IDS incorrectly identifies a benign activity to be malicious while a false negative occurs when the IDS fails to detect a malicious activity [4]. During normal operation, an IDS can generate thousands of false alarms per day [5]. Network intrusion detection systems - no matter if they are anomaly-based or signature-based - share a common problem: the high number of false alerts or false positives. The number of alerts collected by an IDS can be up to 15,000 per day per sensor, and the number of false positives (FP) can be thousands per day. These problems usually cause the final user, the security manager to lose confidence in the alerts, lower the defence levels in order to reduce the number of false positives, or to have an overload of work to recognize true attacks due to IDS mistakes [6].

This paper proposes using machine learning and the k-means data mining algorithm to develop an IDS model with higher efficiency rate and lower false alarms.

1.1 Problem Statement

The study proposes machine learning and the k-means data mining algorithm to develop an IDS model with higher efficiency and lower false using the NSL-KDD data set.

1.2 Research Questions

Consequently, it will answer the following research questions:

1.2.1 To what extent can the k-means detect (i.e. detection rate) attack and normal data?

1.2.2 What are the factors affecting the implementation of an IDS model using k-means data mining algorithm?

1.3 Research Objectives

1.3.1 To be able to detect normal vis-a-vis attack data within the data set.

1.3.2 To be able to identify the false positive rate generated using the k-means algorithm.