Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 83 (2016) 1090 – 1095

Procedia Computer Science

The 6th International Symposium on Frontiers in Ambient and Mobile Systems (FAMS 2016)

Semi-Supervised Statistical Approach for Network Anomaly Detection

Naila Belhadj Aissa[a]*, Mohamed Guerroumi[a]

[a] University of Science and Technology Houari Boumediene, Algiers, Algeria

Abstract

Intrusion Detection Systems (IDS) have become a very important defense measure against security threats. In recent years, computer networks are widely deployed for critical and complex systems, which make them more vulnerable to network attacks. In this paper, we propose a two-stage Semi-supervised Statistical approach for Anomaly Detection (SSAD). The first stage of SSAD aims to build a probabilistic model of normal instances and measures any deviation that exceeds an established threshold. This threshold is deduced from a regularized discriminant function of Maximum Likelihood (ML). The purpose of the second stage is to reduce False Alarm Rate (FAR) through an iterative process that reclassifies anomaly cluster, from the first stage, using a similarity distance and anomaly's cluster dispersion rate. We evaluate the proposed approach on the well-known intrusion detection dataset NSL-KDD and Kyoto 2006+. The experimental results show that SSAD outperforms the Naïve Bayes methods in terms of Detection Rate and False Positive Rate.

Peer-review under responsibility of the Conference Program Chairs

Keywords: Network Security, Anomaly Detection; Maximum Likelihood; chi-square statistic; NSL-KDD dataset ; Kyoto 2006+ dataset

* Corresponding author. Tel.: +213 554 779 601; fax: +0-000-000-0000 .
E-mail address: nbelhadj@usthb.dz

Peer-review under responsibility of the Conference Program Chairs
doi:10.1016/j.procs.2016.04.228

1. Introduction

An intrusion is defined as "any set of actions that attempt to compromise the integrity, confidentiality or availability of a resource"[1]. An Intrusion Detection System (IDS) is a set of components and techniques that aim to monitor network resources or computer activities in order to detect and react to any suspicious action.

IDSs are usually classified into two categories[2, 3]: i) Misuse-based and ii) Anomaly-based. Systems that use misuse-base techniques need a database of known attacks' signature (description) to match against observed activities or audit data. These systems provide very good detection results for specified well-known attacks. However, their inability to detect the unknown attacks is a major drawback. Anomaly-based systems, rely on models of normal behavior of the protected target, any deviation from this model is considered as suspicious. These systems are able to detect previously unknown attacks. However, the number of their false alarms is still high.

Several methods have been proposed to detect anomalies[4], some are based upon computational intelligence (evolutionary computation[5, 6], artificial immune systems[7]…), others use pattern matching or rely on statistical approaches representing normal behavior[8]. These methods can be mainly grouped into three approaches: i) *unsupervised* approach where intrusions are detected with no prior knowledge of a normal/attack behavior. ii) *Supervised* approach where both normal and abnormal behaviors are modeled based on a pre-labeled data tagged as normal or abnormal. iii) In the third *semi-supervised* approach, only normal behavior is modeled using a pre-labeled data. The challenge of this last approach is to find an optimal discriminant function that can efficiently distinguish the target's normal behavior from abnormal instances.

In this paper, we propose a two-stage Semi-supervised Statistical approach for Anomaly Detection (SSAD). In the first stage, we build a probabilistic model based on pre-labeled normal instances, and we classify as anomaly any deviation that exceeds a defined threshold. This threshold is deduced from the Maximum Likelihood (ML) function and the p-value of chi-square (χ^2) distribution. We introduce, for optimization, a variance-covariance matrix's regularization coefficient into the ML function. To reduce the false alarm rate, we propose as second stage, an iterative process based on a similarity distance and the dispersion rate of first stage's output classes.

The remainder of this paper is organized as follows: Section 2 presents some related work on statistical approaches for anomaly detection. Section 3 describes, in details, our proposed approach. Section 4 shows and discusses experimental results. We briefly conclude our work in Section 5.

2. Statistical Approaches for Anomaly Detection

Statistical approaches are commonly based on probabilistic models associated with training data. Several approaches to modeling normal and anomalous data have been applied to intrusion detection: Eleazar Eskin[9] presented a mixture model based technique for detecting anomalies within a dataset that contains a large number of normal instances and relatively few anomalies. His approach uses machine-learning techniques to estimate a probability distribution over this dataset and applies a statistical test to detect anomalies. [10] presents an anomaly detection technique based on a chi-square statistic for intrusion detection into information systems. The authors highlight the great computational cost when using Hotelling's T^2 statistic and discussed the use of a distance measure based on chi-square test statistic instead. Authors in [11] proposed a prototype of a network anomaly detection system, with hierarchical architecture, that uses statistical models and neural network classifiers to detect intrusions. Data are gathered from network traffic, system logs as well as hardware reports. In [12], authors discussed the use of a multivariate statistical analysis of audit trails for host-based intrusion detection. They also discussed and compared the performance of Hotelling's T^2 test and chi-squared test. [13] proposes a statistical processing unit, that utilize application specific knowledge of the network services, to detect anomalous network traffic, specifically the less frequent attacks such as R2L and U2R. A metric is developed which allows the system to automatically search identical characteristics of different service requests. Where authors in [14] used Kolmogorov-Smirnov statistics to model and detect DoS as well as probing attacks. Authors in [15] proposed an approach based on Adaptive Bayesian Algorithm (ABA) for anomaly intrusion detection and compared it with Naïve Bayesian Algorithm using KDD99 dataset. Krueggel and al. [16] suggested an event classification scheme based on Bayesian networks to mitigate two main reasons for large false alarm. According to them, the first reason is the simplistic aggregation of model outputs in the decision phase. The second reason is the lack of integration of additional information into the