Computers & Security (2005) **24**, 160—168

ELSEVIER

Computers & Security

www.elsevier.com/locate/cose

Empirical evaluation of SVM-based masquerade detection using UNIX commands

Han-Sung Kim*, **Sung-Deok Cha**

Division of Computer Science and AITrc/SPIC/IIRTRC, Department of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology (KAIST), 373-1, Guseong-Dong, Yuseong-Gu, Daejeon, Republic of Korea

Received 20 May 2004; revised 27 July 2004; accepted 18 August 2004

KEYWORDS
Intrusion detection;
Masquerade detection;
Anomaly detection;
Machine learning;
Support vector machine (SVM)

Abstract Masqueraders who impersonate other users pose serious threat to computer security. Unfortunately, firewalls or misuse-based intrusion detection systems are generally ineffective in detecting masquerades. Although anomaly detection techniques have long been considered as an effective approach to complement misuse detection techniques, they are not widely used in practice due to poor accuracy and relatively high degree of false alarms. In this paper, we performed an empirical study investigating the effectiveness of SVM (support vector machine) in detecting masquerade activities using two different UNIX command sets used in previous studies [R. Maxion, N. Townsend, Proceedings of international conference on dependable systems and networks (DSN-02), p. 219—28, June 2002; R. Maxion, Proceedings of international conference on dependable systems and networks (DSN-03), p. 5—14, June 2003]. Concept of ''common commands'' was introduced as a feature to more effectively reflect diverse command patterns exhibited by various users. Though still imperfect, we detected masquerades 80.1% and 94.8% of the time, while the previous studies reported the accuracy of 69.3% and 62.8%, respectively, using the same data set containing only the command names. When command names and arguments were included in the experiment, SVM-based approach detected masquerades 87.3% of the time while the previous study, using the same data set, reported 82.1% of accuracy. These combined experiments convincingly demonstrate that SVM is an effective approach to masquerade detection.

* Corresponding author. Tel.: 82 42 869 3575; fax: 82 42 869 8488.
E-mail addresses: kimhs@salmosa.kaist.ac.kr (H.-S. Kim), cha@salmosa.kaist.ac.kr (S.-D. Cha).

0167-4048/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cose.2004.08.007

Introduction

A masquerader is someone who pretends to be another user while invading target user's accounts, directories, or files (Anderson, 1980). The term can be extended to include legitimate but suspicious use of privilege by insiders. While the UNIX root user is usually authorized to read any file, it is highly unusual and even suspicious if for root user to sequentially read all the files stored in user directories or scan the entire file system searching for files containing specific keywords. In such cases, it is prudent to investigate if the root password has been compromised rather than choose to ignore such activities just because the user signed on as the root.

Violation of security policies by insiders is arguably the most serious security threat as the case of FBI agent Hansen (Webster, 2002) illustrates. Such attacks are difficult to detect since commonly deployed security products such as firewalls are unlikely to be effective. Signature-based intrusion detection systems may detect such attacks only if intruders execute the attack codes whose signatures are known. Besides, there are known techniques to evade IDS products (Ptacek, 2002; Rain Forest Puppy, 1999). Therefore, such misuse detection systems can provide only limited help. Access control mechanisms are unlikely to become an effective solution to insider attacks because an inside attacker can install sniffing software to bypass access control mechanisms in place.

Anomaly detection has long been known as an essential component in securing computers and networks. Basic approaches are well established, and techniques such as statistical analysis (Schonlau and DuMouchel, 2001), data mining (Lee and Stolfo, 1998), and various machine learning techniques (Lane, 2000) have been used. Effective anomaly detection mechanisms can defend attacks initiated by either insiders or outsiders. Masquerade detection is a form of anomaly detection because behavior of a masquerade is assumed to be different from that of the real user. Unfortunately, anomaly detection techniques are not widely used in practice primarily because error rate in anomaly detection is still unacceptably high.

In this paper, we demonstrate that SVM (support vector machine), a machine learning technique, is more effective than Naive Bayes technique by repeating the experiments conducted by Maxion and Townsend (2002) and Maxion (2003) using the same data sets in the similar configurations and by comparing the results objectively and quantitatively.

The remainder of the paper is organized as follows: next section introduces the related work on masquerade detection and SVM. Then, we describe architecture for the SVM-based masquerade detection framework, which is followed by the experimental designs and results of two experiments. Conclusions and future works are presented in last section.

Related work

Masquerade detection experiments

Schonlau and DuMouchel (2001) conducted an experiment on masquerade detection using only the command names collected by UNIX `acct` auditing mechanism. They collected 15,000 "truncated" commands each from 70 users and randomly chose 50 of them in the experiment. Commands entered by the rest, 20 users, were used to simulate masquerade activities. Each command set was decomposed into 150 blocks consisting of 100 commands each, and the first 50 blocks, or 5,000 commands, were used as training data and the rest as test data. Experiment administrators randomly inserted 0∼24 command blocks as a means of approximating incursions by masqueraders.

As shown in Table 1, various approaches to masquerade detection were evaluated, and the accuracy of the most effective masquerade detection technique did not exceed 70%.

Maxion and Townsend (2002) conducted another experiment using the same data and same configuration referred to SEA configuration used in Schonlau and DuMouchel (2001). They analyzed data in a different configuration, referred to as the 1 vs 49 configuration, where the first 5,000 commands were used to build a model, and the first 5,000 commands entered by each of the rest of group, 49 users, were used as test data. In SEA configuration, each user's rest 10,000 commands containing simulated masquerade blocks are used as test data. However, 1 vs 49 configuration had a richer source of masquerade data in that 2450 (i.e., 50 blocks $\times$ 49 users) command blocks were used as masquerader data.

Using Naive Bayes classifier, Maxion and Townsend improved the accuracy[1] of masquerader detection from 39.4% to 61.5% while maintaining

[1] In the SEA configuration, total 231 masquerade blocks were inserted in the test data, and the accuracy was calculated, in percents, as follows: accuracy = (the number of blocks detected among masquerading blocks)/231.