Journal of Network and Computer Applications 34 (2011) 1184–1199

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

# Mutual information-based feature selection for intrusion detection systems

Fatemeh Amiri [a,*], MohammadMahdi Rezaei Yousefi [a], Caro Lucas [a], Azadeh Shakery [b], Nasser Yazdani [b]

[a] Center of Excellence, Control and Intelligent Processing, School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran
[b] School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran

ARTICLE INFO

*Article history:*
Received 5 September 2009
Received in revised form
15 December 2010
Accepted 3 January 2011
Available online 14 January 2011

*Keywords:*
Intrusion detection
Least squares support vector machines (LSSVM)
Mutual information (MI)
Linear correlation coefficient
Feature selection algorithm

ABSTRACT

As the network-based technologies become omnipresent, threat detection and prevention for these systems become increasingly important. One of the effective ways to achieve higher security is to use intrusion detection systems, which are software tools used to detect abnormal activities in the computer or network. One technical challenge in intrusion detection systems is the curse of high dimensionality. To overcome this problem, we propose a feature selection phase, which can be generally implemented in any intrusion detection system. In this work, we propose two feature selection algorithms and study the performance of using these algorithms compared to a mutual information-based feature selection method. These feature selection algorithms require the use of a feature goodness measure. We investigate using both a linear and a non-linear measure—*linear correlation coefficient* and *mutual information*, for the feature selection. Further, we introduce an intrusion detection system that uses an improved machine learning based method, Least Squares Support Vector Machine. Experiments on KDD Cup 99 data set address that our proposed *mutual information-based feature selection method* results in detecting intrusions with higher accuracy, especially for *remote to login* (R2L) and *user to remote* (U2R) attacks.



## 1. Introduction

With the rapid progress in the network-based technology and applications, the threat of spammers, attackers and criminal enterprise has also grown accordingly. The 2005 annual computer crime and security survey showed that the total financial losses caused by all kinds of network viruses/intrusions for respondent companies were about US $130 million (C.S. Institute and F.B.O. Investigation, 2005). Furthermore, according to other studies, an average of twenty to forty new vulnerabilities that existed in networking and computer products was detected every month (Patcha and Park, 2007).

The traditional prevention techniques such as user authentication, data encryption, avoiding programming errors and firewalls are used as the first line of defense for computer security (Lazarevic et al., 2003). Lee et al. (2009) proposed a security vulnerability evaluation and patch framework, which enables evaluation of computer program installed on host to detect known vulnerabilities. After evaluation, the vulnerable computer program is patched with the latest patch code. However, intruders can bypass the preventive security tools; thus, a second level of defense is necessary, which is constituted by tools such as *anti-virus software* and *intrusion detection system* (IDS).

Security products like anti-virus softwares have several limitations. They can protect network users from malwares (viruses, Trojan horses, worms and spywares) known with a signature stored in their database. Signature files of many anti-virus products are updated only on a weekly or daily basis. Therefore, computer users are unsafe against new intrusions in the intervals between updates. This is particularly problematic, because new threats can spread across the Internet in a few hours. Furthermore, anti-virus solutions are reactive and do not ensure the safety of the first few computers infected. The signature of threats must first be detected by anti-virus companies, diagnosed and finally a remedy must be deployed. The time of this detection is not predictable.

As opposed to anti-virus programs that detect infected computer programs (Morin and Mé, 2007), an IDS gathers and analyzes information from various areas within a computer or a network (users, processes) in order to identify the subset of activities that violates the security policy. It is designed to give notice that an intruder is trying to get into the system. Traditionally, IDSs have been classified into two categories: *signature-based detection* and *anomaly detection systems*. In *signature-based systems*, attack patterns or behaviors of intruder are modeled and the system will alert once a match is detected. It is able to detect all known attacks with a low

* Corresponding author. Tel.: +98 21 6111 4181; fax: +98 21 88778690.
*E-mail addresses:* fateme.amiri@gmail.com, f.amiri@ece.ut.ac.ir (F. Amiri), rezaei@ece.ut.ac.ir (M. Rezaei Yousefi), lucas@ipm.ir (C. Lucas), shakery@ut.ac.ir (A. Shakery), yazdani@ut.ac.ir (N. Yazdani).

1084-8045/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jnca.2011.01.002