Journal of Network and Computer Applications 34 (2011) 722–730

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

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

Using clustering to improve the KNN-based classifiers for online anomaly network traffic identification

Ming-Yang Su *

Department of Computer Science and Information Engineering, Ming Chuan University, 5 Teh Ming Road, Gwei Shan District, Taoyuan 333, Taiwan, R.O.C.

ARTICLE INFO

Article history:
Received 12 December 2009
Received in revised form
6 September 2010
Accepted 25 October 2010
Available online 4 November 2010

Keywords:
Online anomaly detection
Flooding attacks
DoS (Denial-of-Service) attacks
Genetic algorithm
KNN (*K*-nearest-neighbor) classification
Unsupervised clustering

ABSTRACT

This paper proposes a method to identify flooding attacks in real-time, based on anomaly detection by genetic weighted KNN (*K*-nearest-neighbor) classifiers. A genetic algorithm is used to train an optimal weight vector for features; meanwhile, an unsupervised clustering algorithm is applied to reduce the number of instances in the sampling dataset, in order to shorten training and execution time, as well as to promote the system's overall accuracy. More precisely, instances in the sampling dataset are replaced by less, but more significant, centroids of clusters. According to the proposed method, a system is implemented and evaluated by numerous Denial-of-Service (DoS) attacks. With an embedded weighted KNN classifier, the proposed system could identify a DoS attack from network traffic within a very short time; moreover, the experimental results show that the proposed system could achieve 95.8654% in overall accuracy in the case of 2-fold cross-validation, and 96.25% in overall accuracy for all known attack evaluations. That is, the proposed system possesses both effectiveness and efficiency. Effectiveness is measured by overall accuracy, including detection rate and false alarm rate, and efficiency is measured by the response time during an attack.

1. Introduction

Network intrusion detection systems (NIDSs) are traditionally divided into two broad categories: misuse detection (Lekkas and Mikhailov, 2007; Caswell et al., 2003) and anomaly detection (Toosj and Kahani, 2007; Tsang et al., 2007; Auld et al., 2007). Misuse detection aims to detect known attacks by characterizing the rules that govern these attacks. Thus, rule updates are particularly important and consequently, new definitions are frequently released by NIDS vendors. However, the rapid emergence of new vulnerabilities and exploitations is gradually making misuse detection difficult to trust. Anomaly detection is designed to capture any deviation from the profiles of normal behavior patterns. Anomaly detection is much more suitable than misuse detection for detecting unknown or novel attacks, but it may generate too many false alarms. This paper proposes a system for anomaly detection on DoS attacks by a genetic weighted KNN classifier, which is further enhanced by executing an unsupervised clustering algorithm, named MLBG (Rosenberger and Chehdi, 2000), on the sampling instances in advance.

Most NIDSs emphasize effectiveness but neglect efficiency, especially for anomaly-based NIDSs. Usually, effectiveness is measured by detection rate, false alarm rate, etc., and efficiency is measured by the response time during an attack. Having too many features for an anomaly-based NIDS does not necessarily guarantee good performance, and it certainly delays the detection engine from making a decision. So determining how to select fewer but significant features becomes a vital concern. Furthermore, features should be weighted because their contributions to classification should differ from each other. This study applies a genetic algorithm to weigh all possible features and selects an optimal feature set to construct the proposed real-time NIDS for anomaly network traffic identification. The system performance is measured by weighted KNN classification in which the feature weights react upon distance measurements.

In past studies, some anomaly-based NIDSs focused on the feature weighting and selection, such as Mukkamala and Sung (2002), Sung and Mukkamala (2003), Lee et al. (2006), Abbes et al. (2004), Stein et al. (2005), Hofman et al. (2004), Middlemiss and Dick, (2003), Liao and Vemuri (2002). Mukkamala and Sung (2002) applied the Support Vector Machine (SVM) technique to rank the 41 features provided by KDD CUP99 (The UCI KDD Archive). Sung and Mukkamala also ranked the features by both SVM and neural networks in Sung and Mukkamala (2003). Lee et al. (2006) discussed the feature selections based on a genetic algorithm combined with the Relief Tree, and a genetic algorithm combined with the Naïve Bayesian Network. They also used the KDD CUP99 for an experimental dataset. Abbes et al. (2004) and Stein et al. (2005) both applied decision trees to design their detection systems. Features for tree nodes were selected by a genetic

* Tel.: +886 3 3507001; fax: +886 3 3593874.
E-mail addresses: minysu@mail.mcu.edu.tw, minysu@ms9.hinet.net

1084-8045/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jnca.2010.10.009