Available online at www.sciencedirect.com

ScienceDirect

Pattern Recognition Letters 29 (2008) 918–924

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Design of multiple-level hybrid classifier for intrusion detection system using Bayesian clustering and decision trees

Cheng Xiang *, Png Chin Yong, Lim Swee Meng

Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117576, Singapore

Received 8 May 2007; received in revised form 26 October 2007
Available online 26 January 2008

Communicated by N.S.V. Rao

Abstract

With increasing connectivity between computers, the need to keep networks secure progressively becomes more vital. Intrusion detection systems (IDS) have become an essential component of computer security to supplement existing defenses. This paper proposes a multiple-level hybrid classifier, a novel intrusion detection system, which combines the supervised tree classifiers and unsupervised Bayesian clustering to detect intrusions. Performance of this new approach is measured using the KDDCUP99 dataset and is shown to have high detection and low false alarm rates.

Keywords: Bayesian clustering; Decision tree; False-negative; False-positive; Intrusion detection system (IDS)

1. Introduction

An increasing number of people are now connected to the web as PCs, PDAs and Internet access become more affordable. Current security measures such as firewalls, security policies, and encryption are not sufficient to prevent the compromise of private computers and networks. Therefore, intrusion detection systems (IDS) have become an essential component of computer security to supplement existing defenses. The infamous event of Yahoo being crippled for a few hours from a denial of service (DoS) attack in 2000, had raised lots of concern for network security and generated tremendous interests for designing better IDSs to protect the network.

As manual inspection of the audit trail data generated by operating systems is not feasible due to the incredibly large sizes of the data, data mining has been commonly used to automate the wading through audit data ever since the inception of the field (Denning, 1987; Lunt, 1988). For recent surveys, the reader is referred to Axelsson (2000) and Cabrera and Mehra (2002). Representative designs are Mukkamala et al. (2000) and Lee and Stolfo (2000). In particular, MADAM ID Lee and Stolfo (2000) (for Mining Audit Data for Automated Models for Intrusion Detection) is a good representative of the IDSs built with data mining tools, which has been considered by many researchers in this field as the bench-mark work for intrusion detection systems. However, there are two distinct weaknesses associated with MADAM ID. First of all, the detection rate for DoS attacks is relatively low, (79.9% for known attacks, and 24.3% for unknown attacks), which has to be substantially improved considering the fact that DoS is one of the most notorious and disruptive attacks. Secondly, the framework of MADAM ID is quite complicated, which includes programs for learning classifiers and meta-classifiers (Chan and Stolfo, 1993), association rules (Agrawal et al., 1993) for link analysis, and frequent episodes for sequence analysis (Mannilla et al., 1995).

Several hybrid IDSs have been proposed recently to deal with the complexity of the intrusion detection problem by

* Corresponding author. Fax: +65 67791103.
E-mail address: elexc@nus.edu.sg (C. Xiang).

0167-8655/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2008.01.008

combining different machine learning techniques. Pfahringer (2000), the winner of the KDDCUP99, fused 50 × 10 C5 decision trees using cost-sensitive bagged boosting algorithm, while Levin (2000) used a Kernel Miner to build an optimal decision forest. Both have very low false alarm rates but are unsatisfactory in detecting U2R and R2L attacks. Giacinto et al. (2003) used fusion of multiple classifiers and improved the detection rates for known attacks. However, the detection rates for unknown attacks were significantly lower than those of MADAM ID, which implies that the proposed method may over-fit the training data and not generalize well. Pan et al. (2003) took advantage of different classification abilities of neural networks and the C4.5 decision trees algorithm for different attacks. Depren et al. (2005) suggested a hybrid IDS consisting of an anomaly detection module, a misuse detection module and a decision support system. Zhang and Zulkernine (2006) combined the misuse detection and anomaly detection components in which the random forests algorithm was applied. Most recently, Hwang et al. (2007) proposed a hybrid system combining the advantages of low false-positive rate of signature-based intrusion detection system and the ability of anomaly detection system to detect novel unknown attacks, while Peddabachigari et al. (2007) advocated fusing decision trees and support vector machines as a hierarchical hybrid intelligent system model and an ensemble approach combining the base classifiers.

In order to further increase the intrusion detection rate, as well as to simplify the algorithm, a multiple-level tree classifier was recently proposed in (Xiang et al., 2004) to design an IDS, which contains three-levels of decision tree classification. It was shown to be easy to design and very efficient in detecting known attacks. However, a serious shortcoming of this approach is its high false alarm rate as well as low detection rate for unknown attacks. Thus as an improvement over this design, a new multiple-level hybrid classifier is proposed in this paper to reduce the false alarm rate to an industrially acceptable level while maintaining the low false-negative rate. While MADAM ID and other hybrid IDSs utilized a combination of data mining techniques, the design process of this new approach would be much simpler since only decision trees (Quinlan, 1993) and Bayesian clustering (Cheeseman et al., 1988) are involved.

The rest of the paper is organized as follows. Section 2 gives brief introductions on both decision trees (Quinlan, 1993) and Bayesian clustering (Cheeseman et al., 1988). The structure of the new multiple-level hybrid classifier is described in Section 3. And the experimental results are discussed in Section 4. Section 5 concludes the paper with suggestions for future work.

2. Preliminary

In this section, a brief introduction of the classification algorithms used in the hybrid IDS, i.e., the C4.5 algorithm for building decision trees and the Bayesian clustering algorithm, will be given.

2.1. C4.5 algorithm

The decision tree learning is one of the machine learning approaches for generating classification models. In this paper, C4.5, a later version of the ID3 algorithm (Quinlan, 1993), will be used to construct the decision trees for classification. In ID3, a decision tree is built where each internal node denotes a test on an attribute and each branch represents an outcome of the test. The leaf nodes represent classes or class distributions. The top-most node in a tree is the root node with the highest information gain. After the root node, one of the remaining attribute with the highest information gain is then chosen as the test for the next node. This process continues until all the attributes are compared or when all the samples are all of the same class or there are no remaining attributes on which the samples may be further partitioned.

The attribute with the highest information gain (or greatest entropy reduction) is chosen as the test attribute for the current node. Such an information-theoretic approach minimizes the expected number of tests needed to classify an object and guarantees that a simple (but not necessarily the simplest) tree is found.

Imagine selecting one case at random from a set S of cases and announcing that it belongs to some class C_j. The probability that an arbitrary sample belongs to class C_j is estimated by

$$p_i = \frac{\text{freq}(C_j, S)}{|S|} \qquad (1)$$

where $|S|$ denotes the number of samples in the set S. And so the information it conveys is $-\log_2 p_i$ bits.

Suppose we are given a probability distribution $P = \{p_1, p_2, \ldots, p_n\}$ then the information conveyed by this distribution, also called the entropy of P, is well known as

$$\text{Info}(P) = \sum_{i=1}^{n} -p_i \log_2 p_i \qquad (2)$$

If we partition a set T of samples on the basis of the value of a non-categorical attribute X into sets $T_1, T_2, \ldots, T_m$, then the information needed to identify the class of an element of T becomes the weighted average of the information needed to identify the class of an element of T_i, i.e. the weighted average of $\text{Info}(T_i)$

$$\text{Info}(X, T) = \sum_{i=1}^{m} \frac{|T_i|}{|T|} \times \text{Info}(T_i) \qquad (3)$$

The information gain, $\text{Gain}(X, T)$, is then defined as

$$\text{Gain}(X, T) = \text{Info}(T) - \text{Info}(X, T) \qquad (4)$$

This represents the difference between the information needed to identify an element of T and the information needed to identify an element of T after the value of attribute X has been evaluated. Thus, it is the gain in information due to attribute X.