Chapter 4.48
Risk Factors to Retrieve Anomaly Intrusion Information and Profile User Behavior

Yun Wang
Yale University and Yale-New Haven Health System and Qualidigm, USA

Lee Seidman
Qualidigm, USA

ABSTRACT

The use of network traffic audit data for retrieving anomaly intrusion information and profiling user behavior has been studied previously, but the risk factors associated with attacks remain unclear. This study aimed to identify a set of robust risk factors via the bootstrap resampling and logistic regression modeling methods based on the KDD-cup 1999 data. Of the 46 examined variables, 16 were identified as robust risk factors, and the classification showed similar performances in sensitivity, specificity, and correctly classified rate in comparison with the KDD-cup 1999 winning results that were based on a rule-based decision tree algorithm with all variables. The study emphasizes that the bootstrap simulation and logistic regression modeling techniques offer a novel approach to understanding and identifying risk factors for better information protection on network security.

INTRODUCTION

Statistically based anomaly intrusion detection systems analyze audit trail data to detect anomaly intrusion and profiling use behavior. Although the idea behind intrusion detection is simple (i.e., using normal patterns of legitimate user behavior to identify and distinguish the behavior of an anomalous user) (Anderson, 1972, 1980; Denning, 1987; Helman & Liepins, 1993; Stallings, 2003), abnormal behavior detection is a difficult task to implement because of unpredictable attacks. The ideal intrusion detection system has four goals: (1) to detect a wide variety of intrusions; (2) to detect intrusions in a timely fashion; (3) to present the analysis in a simple format; and (4) to be accurate (Bishop, 2003). Over the past two decades, statistical methods have been used for developing various intrusion detection systems, and achieving these goals has been attempted. Some previously studied methods include, for

example, adaptive detection model (Teng, Chen, & Lu, 1990), principal component analysis (Shyu, Chen, Sarinnapakorn, & Chang, 2003), cluster and multivariate analysis (Taylor & Alves-Foss, 2001; Vaccaro & Liepins, 1989), Hidden Markov Model (Cho & Park, 2003; Gao, Ma, & Yang, 2002), data mining (Anderson, Frivold, & Valdes, 1995; Qu, Vetter, & Jou, 1997; Lee, Stolfo, & Mok, 1999), Bayesian analysis (Barbard, Wu, & Jajodia, 2001), and frequency and simple significance tests (Masum, Ye, Chen, & Noh, 2000; Qin & Hwang, 2004; Ye, Emran, Li, & Chen, 2001; Zhou & Lang, 2003). However, most previous studies have been focused mainly on the first two goals and have been conducted based on the use of all possible variables as independent variables to fit a model. Mukkamala et al. (2003) briefly addressed the data reduction issue, but the knowledge about the degree of significance of an individual variable associated with an attack still remains unclear, and accuracy of such association has not been addressed. A statistical model with a large number of independent variables may not guarantee a high ability of predicting power, and unnecessary variables could cause biases and could lead the model either to overestimate or to underestimate the predicted values. To address these gaps in knowledge, this study, using the bootstrap resample method (Efron & Tibshirani, 1994) and multiple stepwise logistic regression modeling technique (Hosmer & Lemeshow, 2000) sought to identify a small set of risk factors that are robust, statistically significant, and stable to use in detecting anomaly intrusion and profiling user behavior.

METHODS

Data Source

The study sample was drawn from the Third International Knowledge Discovery and Data Mining Tools Competition 1999 data (KDD-cup, 1999), which was created, based on the 1998 Defense Advanced Research Projects Agency (DARPA) Intrusion Detection Evaluation off-line database developed by the Lincoln Laboratory at Massachusetts Institute of Technology (Cunningham et al. 1999). The full KDD-cup data, which included seven weeks of TCP dump network traffic as training data that were processed into about five million connection records, two weeks of testing data, and 34 attack types, were generated on a network that simulated 1,000 Unix hosts and 100 users (Lippmann & Cunningham 2000). The test data do not have the same probability distribution as the training data and include additional specific attack types that were not in the training data. The data unit is a connection that consists of about 100 bytes of information and represents a sequence of TCP packets starting and ending at a fixed time window, between which data flow to and from a source IP address to a destination IP address under predefined protocols. Each connection record is identified as either normal or as a specific attack type. This study used 10% of the training data as a derivation dataset and the full test data as a validation dataset to identify and examine the risk factors.

Outcome and Independent Variables

The outcome of interest was a binary variable that labeled a connection as anomalous (yes/no), which could be any one of the included 38 attack types (24 in the derivation sample and an additional 14 new types in the validation sample). The independent variables included 41 initial variables or features (Stolfo, 2000) across four groups: (1) basic features of individual TCP/IP connections; (2) content features within a connection suggested by domain knowledge; (3) traffic features computed using a two-second time window; and (4) destination features. The type of protocol was categorized into three dummy variables: ICMP (yes/no), TCP (yes/no), and UDP (yes/no); normal