Determining the Minimum Sample Size of Audit Data Required to Profile User Behavior and Detect Anomaly Intrusion

Yun Wang, Center for Outcomes Research and Evaluation, Yale University and Yale New Haven Health, and Qualidigm, USA
Sharon-Lise T. Normand, Department of Biostatistics, Harvard School of Public Health, and Department of Health Care Policy, Harvard Medical School, USA

ABSTRACT

Although statistical modeling techniques have been employed to detect anomaly intrusion and profile user behavior with network traffic data collected from multi-sites (IP addresses), the minimum sample size of audit data required for each site is unclear. Using the Intrusion Detection Evaluation off-line data developed by the Lincoln Laboratory at Massachusetts Institute of Technology under the Defense Advanced Research Projects Agency, this study aimed to address the challenge of determining sample size. Bivariate analysis was employed to construct a composite score to rank each site's probability of being an anomaly, and statistical simulations were conducted to evaluate the ranking variation between the population-based "true" pattern of user behavior and different sample-based "observed" patterns. A sequence of hierarchical random effects logistic regression models was fitted to compare the performance of the full dataset-based and sample-based classifications. The results show that a minimum sample size of 500 per site provides a sensitivity value of 0.85, specificity value of 0.92 and kappa statistic of 0.77. Compared with the full dataset-based model, the minimum sample-based model had a similar Receiver Operating Characteristic area (0.983 vs. 0.997) and a slightly higher misclassification rate (3.16% vs. 1.71%) in detecting abnormal patterns.

Keywords: classification; computer security; intrusion detection; sample size; simulation and modeling IS; user behavior

INTRODUCTION

Over the past decade, network traffic audit data collected from multi-sites (IP addresses) have become increasingly utilized for profiling user behavior in the anomaly intrusion detection area. Statistical methods for the analyses of such data have been studied and have grown substantially, from simple frequency and significant tests (Masum, Ye, Chen, & Noh, 2000; Qin &
Hwang, 2004; Zhou & Lang, 2003) to cluster and multivariable analysis (Taylor & Alves-Foss, 2001; Ye, Emran, Chen, & Vilbert, 2002; Wang, 2005), principal component analysis (Shyu, Chen, Sarinnapakorn & Chang, 2003) and Bayesian analysis (Barbard, Wu, & Jajodia, 2001; Jha, Tan, & Maxion, 2001). However, methods for the collection of such data are relatively limited. Many statistical modeling approaches require accounting for sample size (e.g., number of TCP/IP connections between computers), a factor that has not been well studied in the intrusion detection area. Depending on network traffic, audit data can become very large. For example, 74 hours of logged traffic data from a large enclave network could include as many as 344 million records (24 GB), and 5 hours of tcp-dump packet traces may comprise 110 million connections (Robertson, Siegel, Miller, & Stolfo, 2003). Previous studies have used such large datasets to analyze user behavior and train detection systems, which require more resources to collect and analyze data. While inadequate sample sizes increase both false positive and false negative rates, being able to determine the minimum sample size of audit data required for each site is an essential challenge in the development of a robust and highly efficient intrusion detection model, and has widespread practical application. It can be used, for example, in the mobile and ad hoc computing environment, where only limited information is available for analysis. This study was designed to address this challenge by aiming at three key analytic questions: 1) determining the minimum sample size required per network site to profile user behavior; 2) assessing the impact of different mean values of user behavior on the minimum sample; and 3) evaluating the impact of between-site variance on the minimum sample size. In this study, the terms “site” and “IP address” are interchangeable. They denote a unique network identification that could represent either a signal user or a group of users who share the same IP address (for a dynamic IP address, the range of IP numbers for an index site could be collapsed together to represent the site). In addition, the term “sample size” is equivalent to the “number of TCP/IP connections.” Appendix A provides a glossary of terms.

**METHODS**

**Study Design**

The study was conducted in two stages. The first step involved converting the observed user activities (e.g., average CPU time, log-intimeframe, port of log-in) to a simple composite score for representing user behavior patterns (e.g., normal or abnormal) for each site, and the second step used this score to model the behavior with different sample sizes. Although a variety of statistical modeling approaches can be used to establish such a score, they all require accounting for sample size, the determination of which is unclear at this point. Therefore, a bivariate analysis approach was used to create a composite score ranging from 0 to 1, based on the initial values of variables in the data. The composite score represents the likelihood of being an anomaly-free connection for each TCP/IP connection in the study cohort; for example, a connection with a high score is more likely to be anomaly-free.

The second step required computer simulations using a large volume of network traffic data acquired from each site. For a large number of network trials, the proportion of the number of normal connections...