Online and Adaptive Anomaly Detection: Detecting Intrusions in Unlabelled Audit Data Streams

Wei Wang[*,**], Thomas Guyet[*], René Quiniou[*], Marie-Odile Cordier[*], Florent Masseglia[**]

[*] Projet DREAM, INRIA Rennes/ IRISA, Campus de Beaulieu, 35042 Rennes, France
{thomas.guyet, rene.quiniou, marie-odile.cordier}@irisa.fr
[**]Projet AxIS, INRIA Sophia Antipolis, 2004 route des lucioles - BP 93
06902 Sophia Antipolis, France
wwangemail@gmail.com,florent.masseglia@sophia.inria.fr

1 Introduction

1.1 Issues

Intrusion detection has become a widely studied topic in computer security in recent years. Anomaly detection is an intensive focus in intrusion detection research because of its capability of detecting unknown attacks. Current anomaly IDSs (Intrusion Detection System) have some difficulties for practical use. First, a large amount of precisely labeled data is very difficult to obtain in practical network environments. In contrast, many existing anomaly detection approaches need precisely labeled data to train the detection model. Second, data for intrusion detection is typically steaming and the detection models should be frequently updated with new incoming labeled data. However, many existing anomaly detection methods involve off-line learning, where data is collected, manually labeled and then fed to a learning method to construct normal or attack models. Third, many current anomaly detection approaches assume that the data distribution is stationary and the model is static accordingly. In practice, however, data involved in current network environments evolves continuously. An effective anomaly detection method, therefore, should have adaptive capability to deal with the "concept drift" problem while effectively detects intrusions in unlabelled audit data streams.

1.2 Solution

Our adaptive anomaly intrusion detection method addresses these issues through an online and unsupervised clustering algorithm in data streams, under the assumption that normal data is very large while abnormal data is rare in practical detection environments. Our method adaptively detects attacks with following three steps:

Step 1. Building the initial model with some online clustering algorithms. In this paper we use Affinity Propagation (AP) (Frey and Dueck, 2007) and its extension in streaming environments (Zhang et al., 2008). The first bunch of data is clustered and the exemplars (or cluster centers) as well as their associated items are obtained. Some outliers are identified, marked as *suspicious* and then put into a reservoir.

Step 2. Identifying outliers and updating the model in the streaming environments. As the audit data stream flows in, each incoming data item is compared to the exemplars. If too far