

Chi-square Statistical Profiling for Anomaly Detection

Nong Ye, Qiang Chen and Kyutae Noh
Arizona State University

Purpose

An intrusion is made up of a series of actions to compromise the security of a computer or network system. We develop a multivariate statistical technique based on the chi-square test of goodness-of-fit to detect intrusions. The technique includes two steps: training and testing. In the training, a long-term profile of normal activities in a computer and network system is obtained. In the testing, a signal is produced if an observed event deviates from the norm profile.

Method

Activities in a computer and network system are captured through a stream of audit events. In this study, the training data and the testing data contain computer audit events. There are more than 250 auditable events in a UNIX-based computer system. In this study, we consider 284 types of audit events. Each audit event is represented by the event type. For intrusion detection through anomaly detection, we build a long-term profile of normal activities (norm profile) and compare the observed activities in the recent past with the norm profile to detect an anomaly. The audit events in the recent past from time $t-k$ to the current time t for event type i are represented via a vector of 284 variables, using the exponentially weighted moving average (EWMA) technique as follows.

$X_i(t) = \mathbf{I} \times 1 + (1 - \mathbf{I}) \times X_i(t-1)$ if the observed event at time t falls into the i th event type

$X_i(t) = \mathbf{I} \times 0 + (1 - \mathbf{I}) \times X_i(t-1)$ if the observed event at time t does not fall into the i th event type,

where $i = 1, \dots, 284$. Hence, for each audit event in the training and testing data, we obtain a vector, (X_1, \dots, X_{284}) . By averaging all vectors from the training data, we obtain $\overline{X}_1, \dots, \overline{X}_{284}$ which characterize the long-term profile of normal activities. For a vector of (X_1, \dots, X_{284}) from each audit event in the testing data, a chi-square statistic is computed as follows to measure the discrepancy between the observed activities in the recent past and the long-term norm profile.

$$c^2 = \sum_{i=1}^{284} \frac{(X_i - \overline{X}_i)^2}{\overline{X}_i},$$

If the observed activities in the recent past is compatible with the norm profile, c^2 will be small.

Results

We compute the false alarm rate and the detection rate of the chi-square technique on the testing data. The results show a low false alarm rate and a high detection rate.

New or Breakthrough Aspect of Work

We develop a multivariate statistic based on the chi-square test of goodness-of-fit. Unlike multivariate analysis techniques such as Hotelling's T2 and Bayesian networks, the chi-square multivariate technique does not consider the correlation between variables, which reduces the amount of computation. The testing results in this study show that despite its simplicity, the chi-square multivariate analysis technique still produce good performance in intrusion detection. We also use the EWMA technique to capture and represent computer audit events in the recent past.

Conclusions

The results of this study demonstrate the power of the chi-square multivariate analysis technique as a powerful technique of anomaly detection for intrusion detection.