Review of Anomaly Detection over Noisy Data using Learned Probability Distributions

Author: Eleaszar Eskin, 2000, Reviewed: Wesley Chen for AC299r, Fall 2014

October 2, 2014

1 Backgroud and Motivation

With supervised learning, one of the requirements is that of a clean training set. If there exist anomalies in the training data,

2 Application

Eskin's application was in detecting anomaliest to systems via UNIX system call traces. Anomalies in this case would be intrusions or access by non-desired users. The data set was from DARPA and Stephanie Forrest's group in University of New Mexico containing multiple months of program traces with attacks on certain components.

3 Assumptions

- Normal data can be effectively modelled by any probility distribution
- Anomalous elements are sufficiently distinct from the normal elements
- Anomalies are few (<5%) of the entire data set else model of the normal distribution will get distorted

4 Inital Definitions

- D: generative distribution for entire data
- M: majority distribution
- A: anomalous distribution - a priori assume uniform
- λ: probability of anomalous element (generated from A)
- $(1-\lambda)$: probability of normal element (generated from M)

5 Anomaly Detection Method

Generalized Distribution:

$$D = (1 - \lambda)M + \lambda A$$

Probability Distribution Generation with Function Φ:

$$P_{M_t}(X) = \Phi_M(M_t)(X)$$

$$P_{A_t}(X) = \Phi_A(A_t)(X)$$

Likelihood of Distribution:

$$L_t(D) = \prod_{i=1}^{N} P_D(x_i) = \left((1-\lambda)^{|M_t|} \prod_{x_i \in M_t} P_{M_t}(x_i)\right) \left(\lambda^{|A_t|} \prod_{x_j \in A_t} P_{A_t}(x_j)\right)$$

Log Likelihood:

$$LL_i(D) = |M_t| \log(1-\lambda) + \sum_{x_i \in M_i} \log(P_{M_i}(x_i)) + |A_t| \log(\lambda) + \sum_{x_j \in A_t} \log(P_{A_t}(x_j))$$

Treating x_i as Anomaly:

$$M_t = M_{t-1} \backslash x_t$$

$$A_t = A_{t-1} \cup x_t$$

Score Cutoff:

$$LL_t - LL_{t-1} < c$$

6 Baseline Comparisons

Results under conditions that followed the given assumptions listed above gave anomaly detection rates similar to those of stide and t-stide approaches that required training over clean data sets.