Analyzing Countermeasures to SLT-based Techniques

Marco Barreno
Blaine Nelson
Russell Sears

Motivation

Learning techniques are becoming more widely used in security-sensitive and performance critical applications many of which have significant economic impact.

Relatively little attention has been paid to analyzing the behavior SLT's when the learner is influenced by an attacker.

How much of a threat is an attacker to statistical learning techniques?

- What are the security goals of the application?
- What are the capabilities of the attacker?
- What sort of security properties does the learner have?

SLT Theoretical Bounds

Background

Probably Approximately Correct (PAC) learning - any hypothesis that is consistent with a sufficiently large set of training examples is unlikely to be seriously wrong. Hence PAC learning places bounds on the error of a consistent hypothesis.

Previous Work

Kearns and Li [1993] – Work extending the PAC-learning framework to analyzing SLT algorithms which learn in the presence of malicious noise.

Our Contribution

A categorization of SLT-attacks rather than a general bound and a preliminary analysis of this approach.

Categories of Attacks

- Characteristics of Attacks
  - Does it matter which points are misclassified?
    - Yes – "Specific"
    - No – "Numbing"
  - What sort of errors does the attack cause?
    - Incorrect Acceptance – "Dodging"
    - Incorrect Rejection – "Denial of Service"
  - Does the attack affect learning directly?
    - Yes – "Indoctrination"
    - No – "Analysis"

Novelty Detection

Novelty detection is an important component in many applications where:

- there is an abundance of normal data while abnormal data is scarce.
- even if abnormal data is available, abnormality is not easily characterized.

Novelty Detection Applications

- Fault Detection
- Intrusion Detection
- Virus Detection
- Network Management

Novel Virus Detection

Novelty Detection is a commonly used algorithm for virus detection because:

- Realistic virus behavior is harder to observe than normal behavior.
- Viruses vary widely in their behavior.

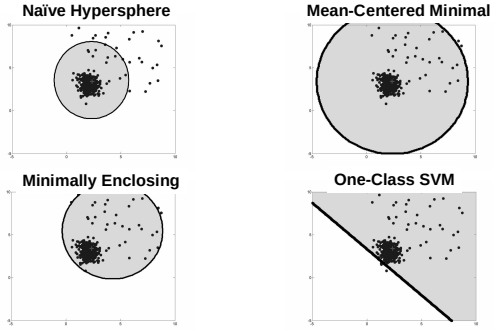
Requirements of an automated virus detector:

- Must adapt to normal changes.
- Must be built with security in mind as virus authors will adapt. E.g. Spam detection.

Concepts in Novelty Detection

- **Decision Boundary** – The boundary in feature space that partitions the space into "normal" and "novel" regions.
- **Kernel Method** – a technique for making linear algorithms nonlinear by implicitly performing the algorithm in a higher dimensional space.
  - Kernel – a function that computes the dot product between two vectors in the higher dimensional space implicitly.
  - All of the novelty detectors considered in this talk are kernelizable.

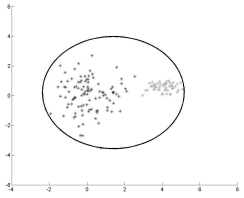
Types of Novelty Detectors



Analysis of Novelty Detectors

Fooling Mean-Centered Approaches

- It is easy for an attacker to expand/shift the region accepted by a mean-centered approach:



Analysis continued

The shift in the mean of a naïve hypersphere:

$$\frac{1}{R}\Delta\overline{X}_{[1,T]} = \frac{1}{R}\sum_{t=1}^{T}\Delta\overline{X}_{[t-1,t]} \le \sum_{i=1}^{T}\frac{\alpha_i}{N+\sum_{j=1}^{i}\alpha_j}$$

$\Delta X_{[i,j]}$ – change in the mean from the i-th to the j-th iteration
α_i – number of points inserted by attacker at the i-th iteration
R – radius of hypersphere
T – number of training iterations
N – initial number of training points

This bound depends on the attack strategy:

- Constant Insertion → Logarithmic shift
- Polynomial Insertion → Logarithmic shift
- Exponential Insertion → Linear shift

Benefits of Bootstrapping

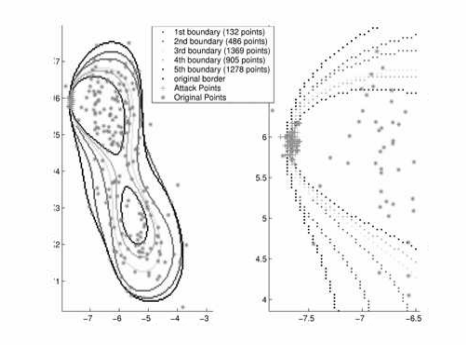
Bootstrapping – a policy for learners that incrementally retrain on new data to only retrain on new data that falls within the current "normal" region.

- A Bootstrapping policy is conservative for a minimal-enclosing hypersphere : Dodging Indoctrination attacks fail.
  - If no outliers are permitted, this property is strict
  - With outliers, the hypersphere can shift, but remains conservative.
- This property extends to One-Class SVMs.

Pitfall of Bootstrapping

- A bootstrapping policy is vulnerable to DOS attacks when outliers are omitted.
- Disturbingly, the hypersphere will become more conservative even when not under attack.
  - Bootstrapping biases the distribution.
  - As new points are added within the "normal" region, the portion of outliers decreases so the hypersphere shrinks.
  - Possible solution: probabilistically accept new points outside of the hypersphere.

Simulated Attack



Future Work

- Much remains on providing defenses against the attacks laid out.
  - While preliminary ideas for defending against SLT attacks have been proposed, a rigorous analysis is needed.
- Multiple attacks with collusion
  - Limiting the control of a single user is an interesting defense mechanism, but doesn't prevent distributed attacks.
- A formal framework is needed to make these concepts rigorous and could provide additional insights.
- Understanding the impact/cost associated with different attacks.

Conclusions

The concept of providing security-analyses for learning applications is essential as such applications are incorporated into security-sensitive environments

We have laid out a basic framework for attacks and applied those attacks to novelty detection providing insight into the strengths and weaknesses of different approaches. We would like the analysis more rigorous by extending the work of Kearns and Li [1993] et. al.

Questions

- A few open questions:
  1. Is there an optimal attack strategy against mean-centered techniques?
  2. Is there a conservative policy for accepting new points that doesn't collapse?
- We welcome any questions, comments, or feedback you may have.