Can Machine Learning Be Secure?

Marco Barreno, Blaine Nelson, Russell Sears, Anthony D. Joseph, J. D. Tygar
Computer Science Division, University of California, Berkeley

The Problem

- **Statistical Learning (SL) is being used increasingly often**
  - IDS, spam filtering, packet routing, etc.
- **However, SL may introduce new vulnerabilities**
  - The SL algorithm itself may be the target of attacks
  - Scant research on this question to date

Setup

- **Learning model**
  - Classification: given *instance* x, predict label y
  - Train learner using a *training set*, and possibly re-train later
  - Training set is labeled with true labels, perhaps with some errors

Attack Model

- **Security Violation**
  - **Integrity** – Intrusion points classified as normal (false negatives)
  - **Availability** – Enough classification errors that learner becomes unusable
  - **Secrecy** – Obtain confidential or private information
- **Influence**
  - **Causative** – Influence over the training data to change learner
  - **Exploratory** – Probing or offline analysis to discover information
- **Specificity (spectrum)**
  - **Targeted** – Focus on a particular point or small set of points
  - **Indiscriminate** – Very general class of points, such as “any false negative”

Abstract

Statistical learning is an invaluable tool that is increasingly being used in security-sensitive applications, but little attention has been paid to the possibility that new vulnerabilities may be introduced by learning systems. We investigate a broad class of statistical learning algorithms and show that their use creates new potential vulnerabilities that an attacker may be able to exploit. We discuss and analyze the range of potential attacks and their effects. We also explore defenses along the lines of adding robustness to the algorithms and selecting appropriate algorithms and model parameters in the first place. Finally, we present some theoretical analysis and experiments to evaluate the attacks and defenses.

Our Approach

- **We focus on a specific choice of attacker:**
  - The attacker's goal is to cause a significant number of errors
  - The attacker can add instances to the training set
  - The attacker can add M instances, has access to the training set, and can arbitrarily choose both label and feature vector for added instances
- **Applications**
  - Intrusion detection in the KDD Cup '99 dataset: Internet connection data, including several attacks
  - Virus detection in the Enron dataset: Corpus of real emails from Enron employees

Defenses

- **Choice of algorithm**
- **Regularization of training data**
  - Removing outliers before training with clustering can reduce potential for attack
- **Adding robustness**
  - Work has been done on robustness against random errors, which may help here even though the assumptions are different

Experiments

- **Purpose**
  - We demonstrate that attacks against the learner present a significant threat to SL systems
  - We also evaluate the potential defenses and ascertain best practices for avoiding vulnerability to attack
- **Experiments are in-progress**

Theoretical work

- **Computational learning theory (PAC learning)**
  - Kearns and Li: the PAC guarantees will be broken if the attacker alters a fraction of the training set greater than the target classification error rate
- **A simple case: hyperspheres for novelty detection**
  - We can bound the movement of the classification boundary for a simple novelty detector
- **Robust statistics**
  - Analyzing the influence of malicious points on the estimators

Future work

- **Completing experiments**
  - Our current project is to complete the experiments and evaluate the relative strengths of the attacks and defenses
- **Other questions remain:**
  - How important is secrecy? How much does the attacker gain by knowing the algorithm used? By having access to the training set? By knowing learned parameters?
  - What information about the algorithm and training data can the attacker extract from interaction with the learner?
  - Can we use game theory to avoid an arms race?

April 2006