Teaching Data-driven Security: A Course on Security Analytics*

Rakesh M. Verma

ReDAS Laboratory

Computer Science Dept.

University of Houston

ReDAS Mission: Research & education in Reasoning, Data Analytics and Security

University of Houston is an NSA/DHS certified center of academic excellence in Information Assurance/Cyber Defense Research AND Education

* Supported by NSF

Outline

- What is Security Analytics?
- Why do we need it?
- A Security Analytics course
- Discussion

What is Security Analytics?

- Adaptation (not direct application) of techniques from
  - Statistics
  - Data Mining
  - Machine Learning
  - Natural Language Processing

  to challenges in cyber security

Why Adaptation?

- Availability/diversity of data
- $$$$(fp), $$(fn)
- 115 phishing emails from July 2012-Aug 2013

[Verma et al., IEEE Security and Privacy, Nov/Dec 2015]

A Security Analytics Course

- Offered at University of Houston in Spring 2015 and 2016
- Senior undergrads (total 22) and grads (total 13)
- Prerequisites:
  - All undergrad math courses for CS major (calculus, prob. & stat., linear algebra, discrete math)
  - Data Structures (3rd CS course)

A Security Analytics Course

- Modular format with a module each on:
  - Quick review of a few key concepts (1 week)
  - Basics of security (4 weeks)
  - Data Mining for security (2.5 weeks)
  - Machine Learning for security (4 weeks)
  - Natural Language Processing for security (3 weeks)

What's in a module?

- Pretest (20-30 min)
- Content
- One homework assignment
- 2-3 Practical exercises
- Posttest (same duration as pretest)
- Quiz (50-55 min)

Basics of Security

- Basic security goals (CIA…) and mechanisms (passwords, crypto, …)
- Malware
- Intrusion detection
- Email security
- Software security

Data Mining for Security

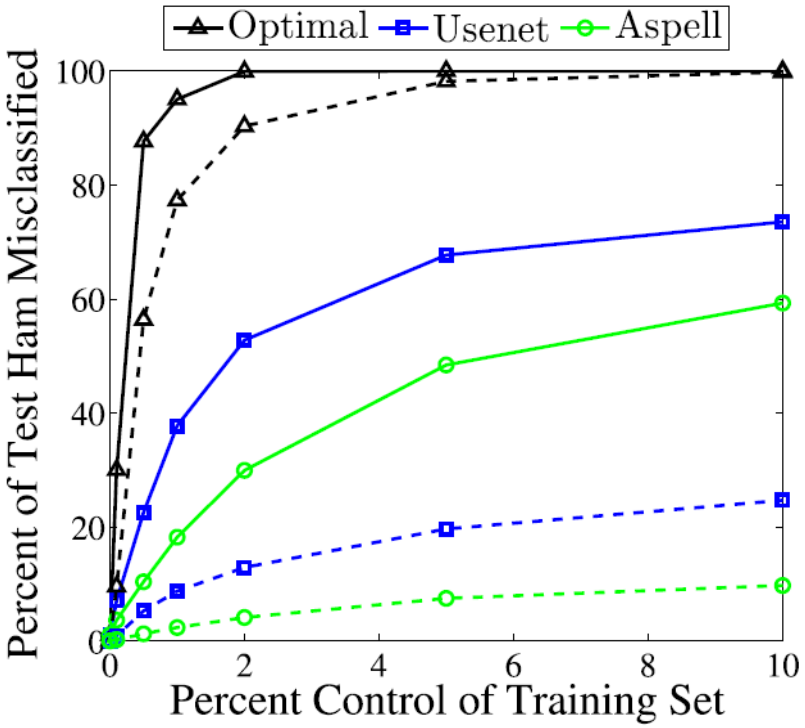
- Data (types and operations)
- Preprocessing and Visualization
- Association rule mining (Computer Crash → DoS)
- Clustering (Virus, Worm, Rootkit)
- Anomaly detection
- Security examples: malware clustering, credit card fraud

Machine Learning for Security

- Basics: kNN, decision tree, naïve bayes
- SVMs (basic and soft margin)
- Neural networks
- One-class learning, semi-supervised learning
- Malware detection, intrusion detection and spam/phishing
- Attack on machine learning models

Some Attacks

- Attacks on SpamBayes
  - Dictionary attack
  - Targeted attack



[A taxonomy of attacks and more examples at:
The Security of Machine Learning by Barreno et al.]

Natural Language Processing for Security

- Language models
- Statistical techniques
- Markov models
- Part-of-speech tagging
- Word sense disambiguation
- Semantics (e.g., via WordNet)
- Applications to security (passwords, spam, phishing, malware)

Grading

- A project is a key component in addition to
  - homework, practical exercises (Weka, R, …), post-test and quizzes
  - Real-world dataset studied with two or three techniques
  - Example: Adapting SVMs and clustering for malware

Feedback

- Quantitative analysis is in progress
- Qualitative feedback examples
  - Security applications need to be more prominent in each model

  + Best course I have taken (undergrad)

  + Learned more about X in this course than I did in courses dedicated to X (grad), where X = data mining, machine learning

Discussion

Thank You!

- Suggestions, Comments and Questions
- But, what about:
  - ensemble learning
  - cost-sensitive learning
  - game theory
  - …..

Contacts

Slides and other materials: http://capex.cs.uh.edu

ReDAS Lab: http://www.cs.uh.edu/~rmverma

Rakesh Verma - rverma@uh.edu