Data Mining Approaches for Intrusion Detection

Serdar Cabuk
Research Assistant
ECE @ Purdue University

Proposed System

- Intrusion Detection in Sensor Networks using Data Mining / Machine Learning Techniques

Intrusion Detection

- Intrusion Prevention is not enough!
- Resources <-> Models <-> Techniques
- **Misuse** vs. **Anomaly Detection**
- What is **Normal**? What is *not* Normal?

Data Mining

“Process of (automatically) extracting models from large stores of data” (Fayyad et al., 1996)

- Classification and / or link and sequence analysis
- STAT511 – Statistical Methods!

Machine Learning

- Concerned with computer programs that automatically improve their performance through experience
- Mining the data results in Machine Learning

DM / ML in ID

- Collect data! -> Data centric method
- Select features!
- Train your machine!
- Extract a pattern / list of patterns!
- Discover the rules!
- **Find the intruder!**

DM / ML in ID (Advanced Issues)

- Adapt changing environment / data!

(i.e. area based labeling in SN)

- Binary labeling vs. Rate based labeling
- Global labeling vs. Local labeling

(i.e. area based labeling in SN)

- Handling *intense* attacks

DM / ML in ID (enough?)

- Noise!
- Evaluation?
- Optimization
  - Feature Selection
  - Sampling
  - Occam’s Razor
- Magic Numbers revisited
- Sensor Network considerations

Example System

"Data Mining Approaches for Intrusion Detection", W. Lee, S. J. Stolfo, 1998

- A simple application of a data mining algorithm, RIPPER, to *sendmail* and *tcpdump*
- Rule generation
- Sliding window

Example System (cont'd)

- Statistical flaws:
  - *How representative is data?*
  - *How long is training phase?*
  - *How representative is training?*
  - *Evaluation – Averaging*
  - *How are outliers injected into system? Does it represent a real-world situation?*
- Had trouble in feature selection

Conclusion

- DM / ML is a well studied discipline
- Large number of DM / ML algorithms are available at free!
- Data analysis may be needed anywhere, including security
- Machine Learning techniques need much more discussion (i.e. Neural Nets)