Scaling Security Analysis

CDT Threats & Risks: Session 9

Matthew Edwards

Dec. 12th, 2019

Section 1

Orientation

Structure: Fortnights

- Threat Modelling
- Unknowns
- Risk Management
- Driving Factors
- Scaling Analysis

Today's Goals

- Overview of security data hurdles.
- Starter code.
- Hands-on challenges.

Section 2

Security Data Analysis

Common machine learning problems

Many cybersecurity problems can be approached as binary or single-class classification problems.

Security as classification

- Is this email spam or ham?
- Is this IP address a botnet C&C server, or not?
- Is this social media account a cyberbully, or not?
- Is this user acting normally, or anomalously?

These problems have a set of existing machine learning solutions (LR, NB, SVM, RF, NN). However, cybersecurity applications are fraught with a set of common issues which make application hard:

1. Gathering representative labelled data
2. Imbalanced classes
3. Concept drift

Labelled data

Labelled data is required for **supervised learning** algorithms.

Each datapoint is described by its *features* and class (or *label*).

By observing associations between combinations of features and the associated labels, a classifier can 'learn' to make predictions about labels based on features.

However, this all presupposes that labelled data is available.

Post-hoc labelling

We can identify attacks/bots/spam/criminals through other means (e.g., manually) and use these as our ‘positive’ labelled items.

Problems

- What about the ‘negative’ items?
- Is our labelling consistent and complete enough to learn from?
- Are we labelling what we think we’re labelling?[1]
- Will this scale?

[1] Wu, X. and Zhang, X., 2016. “Automated inference on criminality using face images.” arXiv preprint arXiv:1611.04135, pp.4038-4052.

Avoid labels

An alternative to supervised learning is **unsupervised** learning – which doesn't require labels!

Problems

- Can only show you patterns & structure in your data.
- The most obvious clusters and patterns in general data are usually unlikely to be related to security.

Semi-supervised learning

A hybrid of the two approaches. Make use of a small number of labelled points to identify classes, but use unsupervised approaches to scale up to the remaining data.

Representative sampling

Malware Experiment

Goal is to build tool to detect malware. So,

- download 100 malware samples from `theZoo`;
- and 100 random 'ordinary files' from somewhere;
- train on 90 malware and 90 ordinary files;
- test on 20 remaining files.
- report accuracy.

Representative sampling

Most security problems are **imbalanced class** problems.

Balanced: 50% malware, 50% goodware.

The real rate of malware appearing in the wild is variable by domain, but a more sane base rate would usually be e.g., 90% goodware, 10% malware.

Some problems are even more imbalanced: 99%, 99.999% goodware.

Problems caused by class imbalance

Evaluating on imbalanced data has implications:

90% ACC:

```python
def classify(datapoint):
  return 'goodware'
```

	Real Positive	Real Negative
Predict Positive	TP	FP
Predict Negative	FN	TN

Calculating accuracy: $\frac{TP+TN}{TP+FP+FN+TN}$

Problems caused by class imbalance

precision $= \frac{TP}{TP+FP}$
recall $= \frac{TP}{TP+FN}$
f-score $= (1+\beta^2) \times \frac{precision \times recall}{(\beta^2 \times precision) + recall}$
(default $\beta = 1$)

More interpretable performance for minority classes. Imbalances can still be problematic for *training* classifiers.

Concept drift

Naive approaches to crossvalidation on longitudinal data can be biased.

Security classifiers often slowly become less performant over time.

Machine learning surprises

Machine learning systems learn to best maximise the goals they are given. Sometimes solutions can turn out to be rather different than researchers intended[2].

Bird & Layzell, 2002 Evolutionary algorithm intended to evolve a timer instead evolves a circuit that picks up the clock from lab PCs.

Murphy 2013 AI trained to play NES games learns to pause the game indefinitely to avoid failing at Tetris.

Chu et al. 2017 CycleGAN, a GAN for converting images into different genres (e.g. horse zebra), gets a reward for transforming images back into their original. It solves this problem by steganographically encoding the original image into the transformed version.

Kelcey 2017 RL trained to maintain simulated car at high speed learns to just spin in circles.

[2] Examples courtesy of `gwern.net/Tanks`

Section 3

Hands-on with spam classification

Section 4

Next Week

Flipped Session

Group 1: Robert, Manolis & Tobias
Group 2: Soo Yee, Priyanka & Hannah

The SpamSlam dataset contains 10,000 labelled examples and 50,000 unlabelled examples. I also have a small dataset which you won't see until next week (for testing). Work in your teams to produce the best classifier you can for the data. Classifiers will be judged by:

1. Performance within the provided labelled data in 10-fold crossvalidation;
2. The best labels for the 50,000 unlabelled examples provided (hint: opportunity for semi-supervised classification)
3. Best performance on brand new data.

Be prepared to discuss the classifiers and features your team tried and settled on.

Request Session

Suggestions:

- Big Data processing
- Web scraping
- Free time
- ...