IntrusionDetection.Systems

ECE 849

Shahzad Ali

http://www.cs.cmu.edu/~cheeko/intrusion

myTaxonomy.IDS

One way to look at IDS is to look at the Machine Learning (ML) Technique used to build them.

- Supervised Learning: Show them normal and they detect abnormalities.
  - Pattern Matching, Signature based, **Data-Mining**, Genetic Algorithms, **Genetic Programming**.
- Unsupervised Learning: Show them normal and some abnormal and they detect other abnormalities
  - Neural Networks, Hidden Markov Models, **Frequentist**.

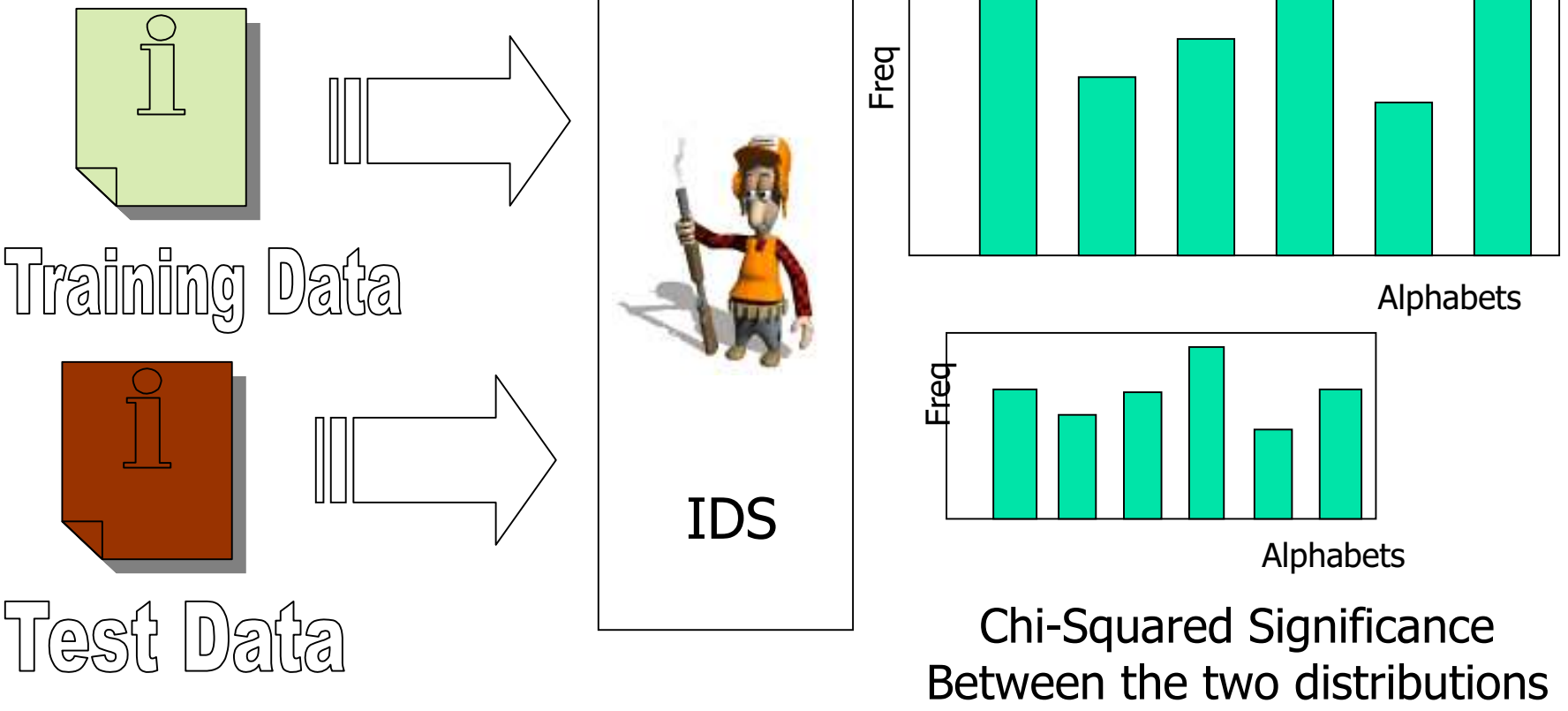
Data.sets

- Synthetic Data: not synthesized by me.
- Alphabet size 6, anomaly size 3
- 11 sets of training/test data.
- Training data is a stream of 0.26M characters in varying randomness.
- Test data is a stream of 0.26M characters with 160 anomalies.
- There is a key file with each test file indicating where the anomalies are.

IDS->how.to.test

Supervised	Unsupervised
▪ Train IDS on training data + 20% of test-data (which 20%??) ▪ Test IDS on 80% of test data. ▪ Compare its alarms with the actual anomalies. ▪ Change Parameters and repeat. ▪ Evaluate.	▪ Train IDS on training data ▪ Test IDS on 100% of the test data. ▪ Compare its alarms with the actual anomalies. ▪ Change Parameters and repeat. ▪ Evaluate.

IDS.Frequentist



IDS.Data-mining

- Use a rule induction tool such as RIPPER (Cohen95) to induce rules about the data. (Lee, Stolfo)
- We used another tool, CN2.
- Reformatted the data-sets to be in sets of size n, where 3<n<13.

  A B C D E C D E F A A C E F D … => A B C D E n;
  C D E F A n;
  A C E F D y;

- CN2 = Decision Trees + Laplace Error Estimate = ordered/unordered rule set.

IDS.GP (Genetic Programming)

- GP is not GA!!
- GA represents solution as string of bits.
- GP represents solution as a tree.
- Operators: +,-,*,/,max,min,sqrt.
- Terminals: 1..n, where n is the size of the pattern so that terminal i means the i^{th} symbol in the pattern.
- GP Parameter: After analysis (trust me)

  Population = 50, Generations = 20,

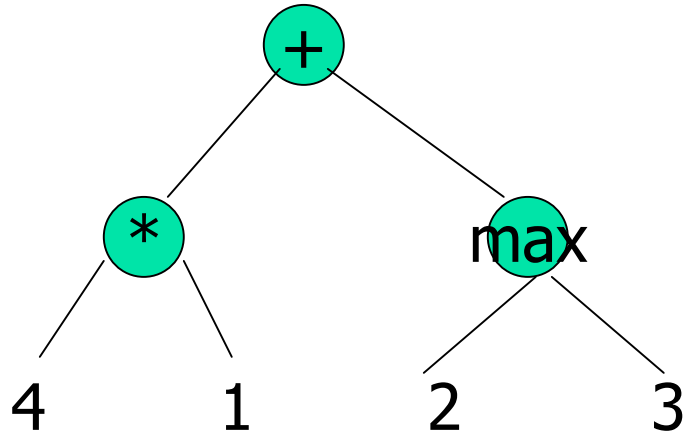
  Patterns = 3,6,9,12

IDS.GP2 (Genetic Programming)

- Translate symbols into numbers.

A=-3,B=-2,C=-1,D=1,E=2,F=3

- What one GP of pattern size 4 might look like?
- What is the value of the GP?
- Decision: If value < 0, anomaly, else normal (arbitrary choice).



A C D E => E*A + max(C,D)

2*(-3) + max(-1,1) = -5

ANOMALY!

IDS.GP3 (Genetic Programming)

- Evaluation: Run the GP over the training set + 20 % of test set.
- Score using formula:

  Fitness = false_alarms + 1000*misses

IDS.GP4 (Genetic Programming)

- Start with a random population of GP's.
- Evaluate them using the fitness function
- Choose the best.
- Evolve next generation using crossovers + mutations.
- The best GP of each generation is also run over the test data for evaluation.
- Result is using the best GP of the last generation.

IDS.Conclusions

- Tried three different IDS schemes.
- Evaluated performance of detectors.
- Frequentist only works for *easy* cases.
- Data-mining is too conservative, but accurate!
- Used a novel approach for IDS; GP.
- Promising results for GP.

Need to thank Roy Maxion and his students for help.