

GP based Intrusion Detection

An Example of Training GP on Large Data Sets

March 2k3 CSCI6506 1

Problem Context

- Large Data Set
- Transparent Solutions
- High Throughput
- Minimum Number of Input Features
- Binary Classification
 - attack/ normal
- Require Labeled data set
- 1999 Knowledge and Data-mining Competition
- TCP dump converted into 'Connection' information
 - Preprocessing with a network analyzer

March 2k3 CSCI6506 2

Connection Features

- Basic Features
 - Properties of TCP packet
- Content Features
 - Pay load features
- Time-based Traffic Features
 - Statistics over 2 second window
- Host-based Traffic Features
 - Statistics over last 100 connections

March 2k3 CSCI6506 3

Connection Types

| Type | Training | Test |
|-------------------|----------|--------|
| Normal | 19.69% | 19.48% |
| Probe | 0.83% | 1.34% |
| Denial of Service | 79.24% | 73.9% |
| User to Root | 0.01% | 0.07% |
| Remote to Local | 0.23% | 5.2% |

March 2k3 CSCI6506 4

Content Per Data Set

| Connection | Training | Test |
|------------|----------|---------|
| Normal | 97,249 | 60,577 |
| Attack | 396,744 | 250,424 |

March 2k3

CSCI6506

5

GP Design Decisions

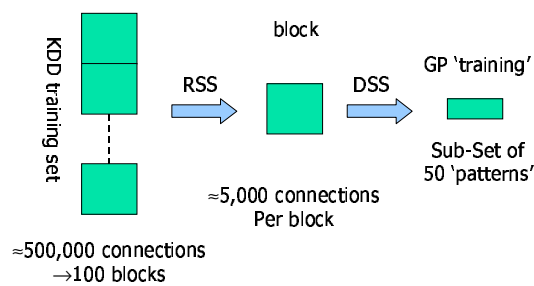
- *Cannot* train over *entire* data set *directly*
- Features Employed
 - Representation of Time
- Functional Set
- Generic GP Parameter Selection

March 2k3

CSCI6506

6

Data Sub-Sampling – Architecture



March 2k3

CSCI6506

7

Dynamic Subset Selection

- Level 1
 - Single block resides within RAM
 - Random Selection of block
- Level 2
 - How many GP 'iterations' per block?
 - Composition of Sub-Set?
 - Rate of sub-set refresh?

March 2k3

CSCI6506

8

GP 'iterations' per block

- Let 1 iteration = 6 tournaments
- Iterations per block =
 - (Max Sub-Set Selections)
 - ×
 - (Miss Classification rate of best individual on previous instance of block)

March 2k3 CSCI6506 9

Max Sub-Set Selections?

March 2k3 CSCI6506 10

Connection Features Employed

- 1st eight (of nine) 'basic' TCP 'features'
 - Duration
 - Protocol
 - Service
 - Normal or Error
 - Number data bytes from source to dest.
 - Number data bytes from dest. to source
 - Number of wrong fragments

March 2k3 CSCI6506 11

Representation of Time

- Let 'x(s)' denote connection at sequence position 's'.
 - x = any of the 8 'basic TCP features'
 - s = {0, 4, 8, 12, 16, 20, 24, 32}
- GP may index history relative to sample 's', or
 - [x(s), x(s - 4), ..., x(s - 32)]

March 2k3 CSCI6506 12

Linear-GP Parameters

| Parameter | Value | Parameter | Value |
|-------------|-------|-----------|---|
| Pop. Size | 125 | Function | {%,*,+,-} |
| Max. Instr. | 256 | Terminal | {I ₀ ...I ₆₃ , 0...255} |
| P(Xover) | 0.9 | RSS | ≈5,000 |
| P(mutate) | 0.5 | DSS | 50 |
| P(Swap) | 0.9 | RSS iter. | 1,000 |
| Tournament | 4 | DSS iter. | 100 |
| # Register | 8 | Wrapper | 0 if R0 ≤ 0, else 1 |

March 2k3

CSCI6506

13

Evaluation

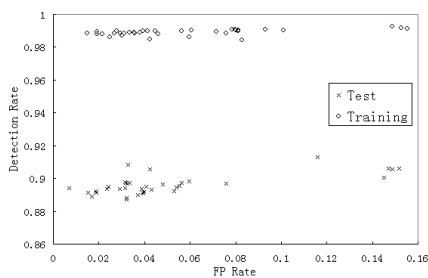
- Platform
 - 1 GHz Pentium III; 256 Mbyte RAM.
 - Each run ≈15 minutes, or 10 hr for 40 runs;
- Detection Rate = $1 - \frac{\text{\#False Negatives}}{\text{Total Attacks}}$
- False Positive Rate = $\frac{\text{\#False Positives}}{\text{Total Normal}}$

March 2k3

CSCI6506

14

FP & Detection rate over 40 runs



March 2k3

CSCI6506

15

GP versus Competition Best – Test Data Alone

| Parameter | Detection | FP Rate |
|-----------------------------|-----------|---------|
| 1 st Place Entry | 90.88% | 0.447% |
| 2 nd Place Entry | 91.525% | 0.576% |
| GP – best FP | 89.4% | 0.682% |
| GP – best detection | 90.825% | 3.27% |

March 2k3

CSCI6506

16

Test Performance Snap shot – 'Seen' attack types

| Connection Type | %Missclass. | Total Examples |
|-----------------|-------------|----------------|
| Neptune | 0 | 58,001 |
| Portswep | 0 | 254 |
| Land | 0 | 9 |
| Nmap | 0 | 84 |
| Smurf | 0.077 | 164,091 |
| Satan | 3.552 | 1,633 |
| Normal | 3.267 | 60,577 |

March 2k3

CSCI6506

17

Test Performance Snap shot – 'Unseen' attack types

| Connection Type | %Missclass. | Total Examples |
|-----------------|-------------|----------------|
| Udpstorm | 0 | 2 |
| Prostable | 3.0 | 759 |
| Saint | 5.98 | 736 |
| Mscan | 8.45 | 1,053 |
| Httpunnel | 15.8 | 158 |
| Phf | 50 | 2 |
| Apache2 | 65.5 | 794 |

March 2k3

CSCI6506

18

Solution Transparency

$$\frac{(20 - I_a) \times I_b}{I_c} - I_a - I_d$$

$$\frac{I_c}{I_c}$$

March 2k3

CSCI6506

19

Conclusion

- Efficient Learning on large dataset
- Concise Rule
- Low a priori knowledge
- Future
 - Different Cost Function
 - Entire Set of Features

March 2k3

CSCI6506

20



References/ further reading

1. Song D., Heywood M.I., Zincir-Heywood A.N., A Linear Genetic Programming Approach to Intrusion Detection, to appear in GECCO'2003.
2. Elkan C.: (2000) Results of the KDD'99 Classifier Learning Contest. SIGKDD Explorations. ACM SIGKDD. 1(2) pp 63-64.
3. Gathercole C., Ross P.: (1994) Dynamic Training Subset Selection for Supervised Learning in Genetic Programming. Parallel Problem Solving From Nature III. LNCS 866, pp 312-321.