Intrusion Detection with Genetic Algorithms and Fuzzy Logic

Emma Ireland

Division of Science and Mathematics
University of Minnesota, Morris
Morris, Minnesota, USA

December 2013
UMM CSci Senior Seminar Conference

The Big Picture

-
- have story here,
- also explain root somewhere
-
-

Outline

Outline

Types of Networking Attacks

- Denial of Service (DoS): attacker makes a machine inaccessible to a user by making it too busy to serve legitimate requests.
- Remote to User (R2L): attacker tries to gain access to things a local user would have on the machine.
- User to Root (U2R): attacker starts out with access on the machine and then tries to gain root access to the system.
- Probe: attacker examines a machine in order to collect information about weaknesses or vulnerabilities that in the future could be used to compromise the system.

Detection Methodologies

- Signature-based detection: compares well-known patterns of attacks that are already known to the IDS against captured events in order to identify a possible attack.
  - It is a simple and effective way to detect known attacks, but is ineffective against new kinds of unknown attacks.
- Anomaly-based detection: looks for patterns of activity that are rare and uncommon.
  - It is harder to do than signature-based detection, but it can be an effective way to detect new, unknown attacks.

Rules

- A commonly used approach for detecting intrusions is to use rules.
- If-Then format: If (*condition*) then (*consequence*).
  - The condition is composed of one or more features, and the consequence says if it is an intrusion or not.
  - If *duration = 4* then *intrusion*.

KDD99

- Generated by simulating a military network environment in 1999.
- Has long been a standard data set for intrusion detection.
- Data was processed into 5 million records.
  - A record is a sequence of TCP packets, between which data flows to and from a source IP address to a target IP address.
- Data in the set is classified as normal or attack activity.
- Uses 41 features, which are properties of a record that are used to describe the activity and help to distinguish normal connections from attacks.

Some Features of KDD99

- duration: length of the record in seconds. in seconds.
- num_failed_logins: number of failed login attempts.
- root_shell: returns 1 if root shell is obtained, else returns 0.

RLD09

- RLD09 was created because KDD99 is 14 years old, and newer attack types are not included in it because of its age.
- Data was captured from a university in Bangkok, Thailand.
- 17 different types of attacks (divided into denial of service and probe attacks).

- .....maybe have something like "the authors of [6] used only the numerical features of KDD99 (34 out of 41)." if I decide to explain that paper.

Training and Testing Sets

- A common technique is to divide the data set into two subsets, a *training set* and a *testing set*.
- The given algorithm is then trained on the training set to look for patterns.
- These patterns are then verified using the test set.

Determining the Accuracy of an Algorithm

		Predicted: Not Attack	Predicted: Attack
Actual	Not Attack	True Negative (TN)	False Positive (FP)
	Attack	False Negative (FN)	True Positive (FP)

- Detection rate (DR): the percentage of normal and attack activity correctly classified from the total number of data records.

- .....maybe talk about the other paper's DR?

Genetic Algorithms

- GAs are a search technique used to find solutions to problems. Operations analogous to biological mutation, selection, and crossover are used to evolve and improve solutions.
- Possible solutions to problems can be represented in a variety of problem dependent ways, such as bit strings.
- First, a randomly generated population of potential solutions is created. Then mutation, crossover, and selection are applied to each generation until an acceptable solution is found or some time limit is exceeded.
- Mutation: random bits in an individual, or possible solution, are randomly changed.
- Crossover: Two individuals swap sequences of bits to form two new individuals.
- Selection: individuals that have better fitness are chosen to be parents.
- The fitness of an individual is specified by the fitness function.

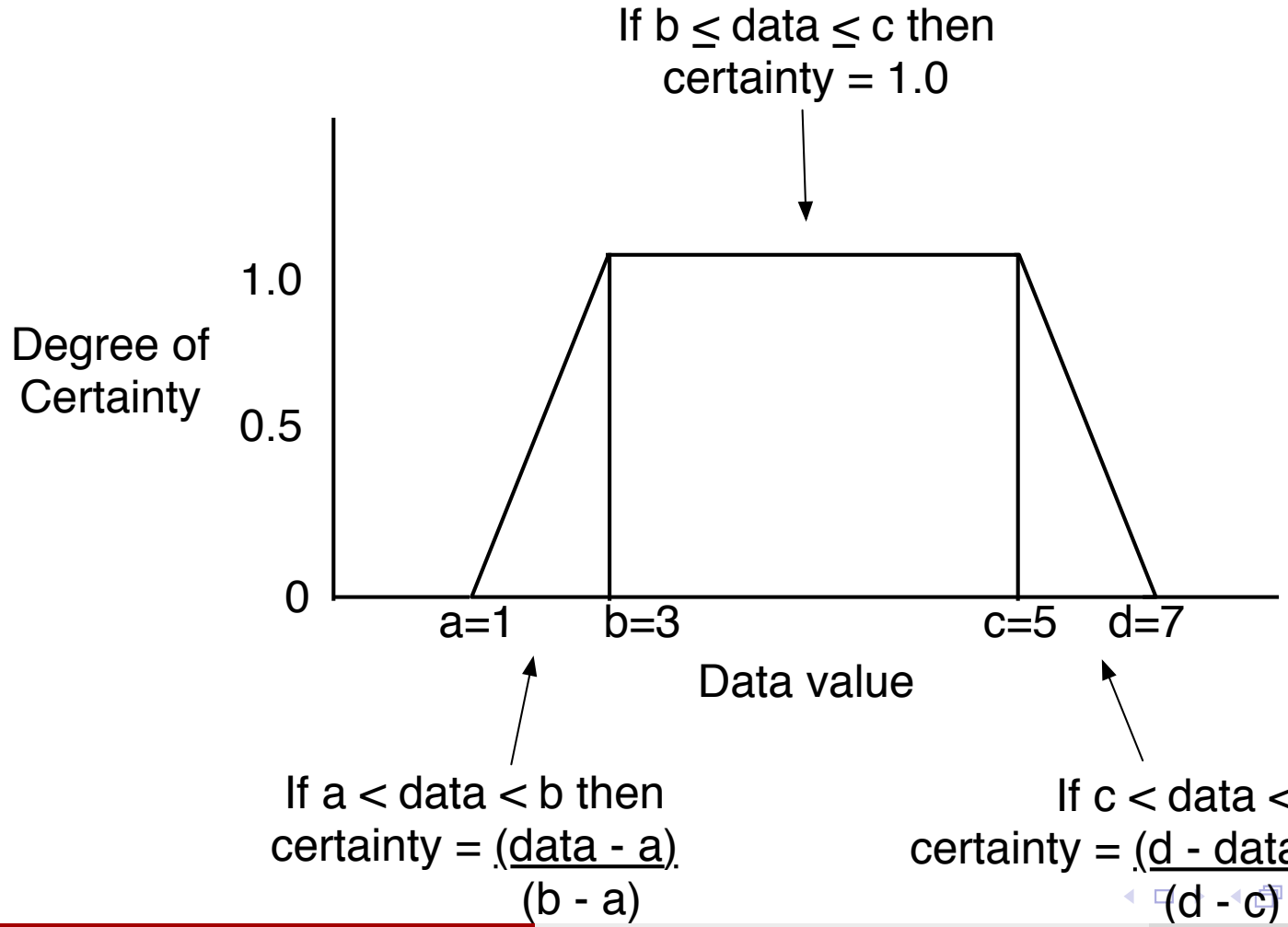
Outline

Fuzzy Logic

- Attacks on systems do not always have a fixed pattern, so fuzzy logic is used to detect patterns that have a behavior that is between normal and unusual.
- Fuzzy logic rules are similar to the rules described before, except that consequence is a certainty factor.
  - If (*duration* = 6) then (*the degree of certainty of it being an attack is 0.5*).

Finding the Degree of Certainty of a Record Being an Attack

Suppose that the feature is duration, and it is 6 seconds. Then data=6.



Encoding of Features and Rules

- The four parameters are encoded into blocks. Each block is a feature with values between 0 and 7.

010	011	100	101
a=2	b=3	c=4	d=5

- A rule has one block for each of 12 features followed at the end by a marker indicating the type of attack.

010	011	100	101	.....	010	011	101	111	DoS
a=2	b=3	c=4	d=5	.....	a=2	b=3	c=5	d=7	
		Block 1				Block 12			Type

Figure: Based on [Jongsuebsuk-Wattanapongsakorn-Charnsripinyo, 2013]

- The degree of certainty is computed for each of the 12 blocks, and if the sum of those is greater than a threshold, then it will be declared as an attack.

Algorithm Overview

- The algorithm generates rules and improves rules.
- One record is passed into a rule.
- Each feature in a record is matched to one block of the rule.
- The parameters of each block measure the degree of certainty of an attack using the trapezoidal fuzzy rule shape.
- The sum of the degrees of certainty from each block are then compared with a threshold to determine if the record represents an attack or normal behavior.

Algorithm

```
for each rule do
    for each record do
        for each feature do
            prob = fuzzy();
            totalprob = totalprob + prob;
        end for
        if totalprob > threshold then
            class is attack;
        else
            class is normal;
        end if
    end for
    find A, B, α, and β
end for
calculate fitness
crossover(), mutation()
```

The fitness function in the algorithm is:

$$\frac{\alpha}{A} - \frac{\beta}{B}$$

A: # of attack records.
B: # of normal records.
α: # of attack records correctly identified as attack.
β: # of normal records incorrectly classified as attack.

Experiments

- A variety of experiments were run. Two experiments used just RLD09. Three experiments used both the RLD09 and KDD99 in order to compare how the fuzzy GA would perform on both.

Experiments Using Only RLD09

Experiment 1

- Fuzzy GA was used to create DoS and probe detection rules. Both of the DoS and probe rules were then used together in the testing process to identify attacks from the testing data set.
- 10,000 records were used for the training set, 26,500 records were used for the test set.

- The following algorithm was used to identify attacks and normal activity in this experiment.

if dos_rule = yes or probe_rule = yes **then**
  This record is an attack;
else
  This record is normal;
end if

Experiments Using Only RLD09

Experiment 1 Results

	Attack	Normal	Total	FP(%)	FN(%)	DR(%)
DoS Training	1499	8501	10000	1.46	47.50	91.64
Probe Training	2496	7504	10000	1.83	15.38	94.79
Testing	10500	16000	26500	1.13	4.10	97.92

Experiments Using Only RLD09

Experiment 2

- Attacks were pulled out of the training set and kept for unknown data testing. This was to test that the fuzzy GA could detect unknown attacks.
- Used fuzzy GA and a decision tree algorithm.
- For each test case there were 13 attack types plus normal activity that were in the training data set. Three attack types were used for the unknown testing data set.

Experiments Using Only RLD09

Experiment 2 Results (7 tests were run in total, 5 are shown here.)

Test Case	Unknown Attacks	Decision Tree DR (%)	Fuzzy Genetic DR (%)
1	Adv Port Scan (Probe) Ack Scan (Probe) Xmas Tree (Probe)	Avg = 98.33	Avg = 100
2	UDP Flood (DoS) Host Scan (Probe) UDP Scan (Probe)	Avg = 46.65	Avg = 99.80
3	Jping (DoS) Syn Scan (Probe) Fin Scan (Probe)	Avg = 99.70	Avg = 98.75
4	UDP Flood (DoS) RCP Scan (Probe) Fin Scan (Probe)	Avg = 70.35	Avg = 98.15
5	Http Flood (DoS) RCP Scan (Probe) Fin Scan (Probe)	Avg = 99.94	Avg = 97.50

Experiments Using Both RLD09 and KDD99

Experiment 1

- Used fuzzy GA to classify normal activity and attacks from KDD99 and RLD09.

Data set	Attack	Normal	FP (%)	FN (%)	DR (%)
KDD99	160,117	39,337	0.13	1.55	98.72
RLD09	10,500	16,000	1.14	3.39	97.97

Experiments Using Both RLD09 and KDD99

Experiment 2

- Used the fuzzy GA to classify types of attacks in KDD99.
- 158,597 records of DoS attacks. 1,500 records of probe attacks.
- 10 tests were run in total, 5 are shown here.

Test	Attack	Type	FP (%)	FN (%)	DR (%)
1	Back	DoS	85.33	0.00	16.56
2	Smurf	DoS	0.76	0.10	99.73
3	Neptune	DoS	0.15	0.34	99.75
4	Portsweep	Probe	6.40	0.00	93.66
5	Satan	Probe	0.74	3.75	99.22

Experiments Using Both RLD09 and KDD99

Experiment 3

- Used the fuzzy GA to classify types of attacks in RLD09.
- 6,400 records of DoS attacks. 10,400 records of probe attacks.
- 17 tests were run in total, 6 are shown here.

Test	Attack	Type	FP (%)	FN (%)	DR (%)
1	HTTP Flood	DoS	0.36	3.5	99.46
2	Smurf	DoS	0.02	0	99.98
3	UDP Flood	DoS	11.06	0	89.59
4	Fin Scan	Probe	2.58	0	97.50
5	IP Scan	Probe	13.01	16.4	86.89
6	Syn Scan	Probe	0.65	4.2	99.24

Outline

Conclusions

- The fuzzy genetic algorithm had a higher detection rate than a decision tree algorithm in most cases.
- Fuzzy genetic algorithms are good at detecting unknown attacks.
- The use of fuzzy genetic algorithms in intrusion detection is an effective way of detecting attacks.

Thanks!

Thank you for your time and attention!

Questions?

References

Jongsuebsuk, P. and Wattanapongsakorn, N. and Charnsripinyo, C.
Network intrusion detection with Fuzzy Genetic Algorithm for unknown attacks.
In *2013 International Conference on Information Networking (ICOIN)*, pages 1-5, 2013.

Jongsuebsuk, P. and Wattanapongsakorn, N. and Charnsripinyo, C.
Real-time intrusion detection with fuzzy genetic algorithm.
In *2013 10th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pages 1-6, 2013.

Mohammad Sazzadul Hoque, Md. Abdul Mukit, Md. Abu Naser Bikas.
An Implementation of Intrusion Detection System Using Genetic Algorithm.
CoRR, abs/1204.1336, 2012.

See my Senior Seminar paper for additional references.