IEEE NFV-SDN 2016

PREDICTING NETWORK ATTACK PATTERNS IN SDN USING MACHINE LEARNING APPROACH

SAURAV NANDA,

FAHEEM ZAFARI, CASIMER DECUSATIS,

ERIC WEDAA AND BAIJIAN YANG

PURDUE
COLLEGE OF TECHNOLOGY

COMPUTER AND
INFORMATION TECHNOLOGY

Agenda

❑ Motivation

❑ Introduction

❑ Problem Statement

❑ Methodology

❑ Experimentation

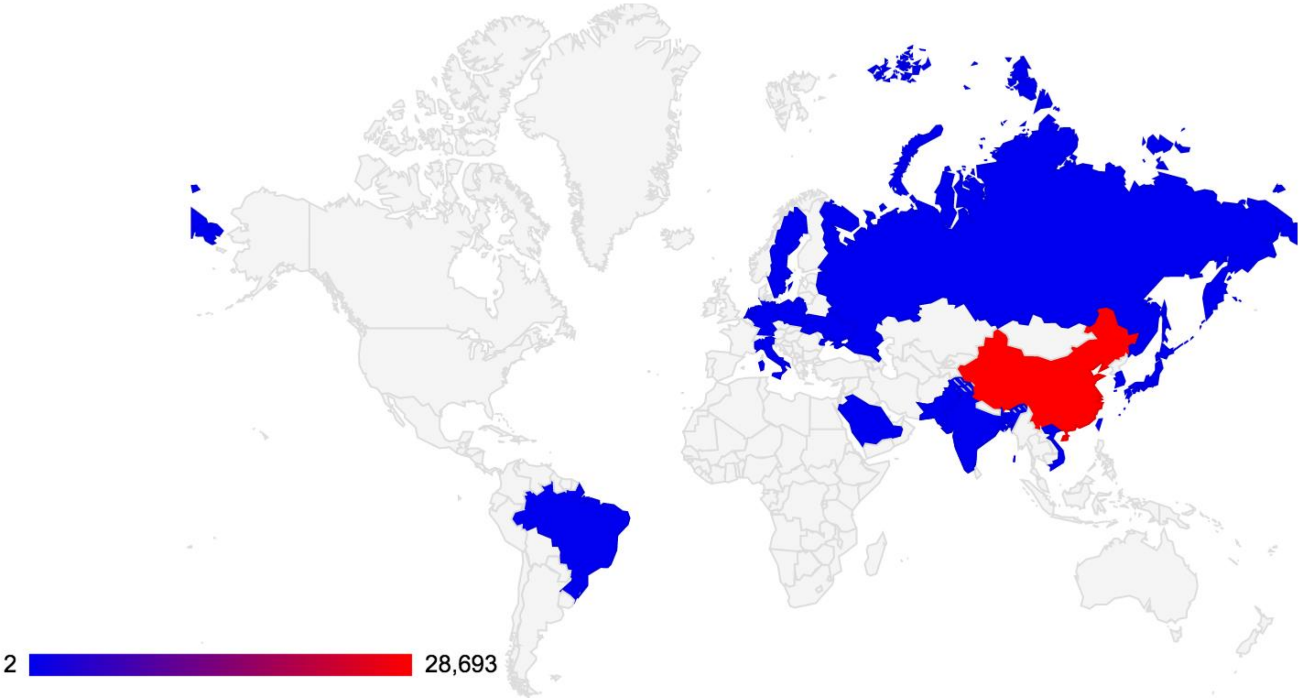
❑ Results

❑ References

Motivation

- **LongTail Project**
  - Analyzes SSH brute force attacks
  - Statistically quantifies:
    - IP addresses
    - Accounts
    - Passwords
    - Account/password pairs
    - Analyzing attack patterns

Log Analysis, All SSH Ports



Data on this page last updated on	Tue Nov 8 00:27:05 EST 2016
Number of ssh login attempts today	29503
Number of usernames seen today	61
Number of unique usernames seen today	61
Number of passwords seen today	8121
Number of unique passwords seen today	6
Number of IP addresses seen today	43
Number of unique IP addresses seen today	7

http://longtail.it.marist.edu/honey/historical/2016/11/07/

Sample Data from LongTail

IP	Lifetime In Days	Botnet	First Date Seen	Last Date Seen	## of Attack Patterns Recorded
49.236.204.180	640.23	pink_roses	2015/01/12 13:26:34	2016/10/13 19:53:54	272
122.160.154.221	639.52	big_botnet	2015/02/02 16:46:25	2016/11/03 06:20:11	11
218.65.30.92	636.49	friends_of_sshPsycho_IP_addresses	2015/02/03 22:33:28	2016/11/01 11:23:40	556
59.51.24.186	620.73	pink_roses	2015/02/06 16:47:43	2016/10/19 11:26:04	124
222.186.56.42	618.37	15-07-01-botnet-20	2015/02/07 07:45:01	2016/10/17 17:37:26	12

http://longtail.it.marist.edu/honey/current_attackers_lifespan.shtml

Log Analysis of IP Attacks

100.38.47.218
1 lines, dict-e53664bda267cedce5900c80d1902ef2.txt To: edub Attack #: 1 started on 2016/03/24 13:49:33
1 lines, dict-3d520cba13d3e60f92b8d6874428e82f.txt To: edu_c Attack #: 2 started on 2016/04/07 08:14:54

100.38.74.99
1 lines, dict-02a719d9d242acd4fcd8cc6da9f6cfbd.txt To: shepherd Attack #: 1 started on 2016/07/18 12:31:34

101.0.44.181
1 lines, dict-d64b8ef614272f5c703f4ae0cf1c51d7.txt To: syrtest Attack #: 1 started on 2015/08/25 00:24:02

101.0.44.231
1 lines, dict-ad6234f04947b500af48eba5d7f4a6fd.txt To: kippo2Jul Attack #: 1 started on 2015/07/24 11:44:45

101.0.44.236
1 lines, dict-864992c0102f84225736a7291a3791eb.txt To: edu_c Attack #: 1 started on 2015/07/31 16:08:51

http://longtail.it.marist.edu/honey/ip_attacks.shtml#58.206.126.29

Problem Statement

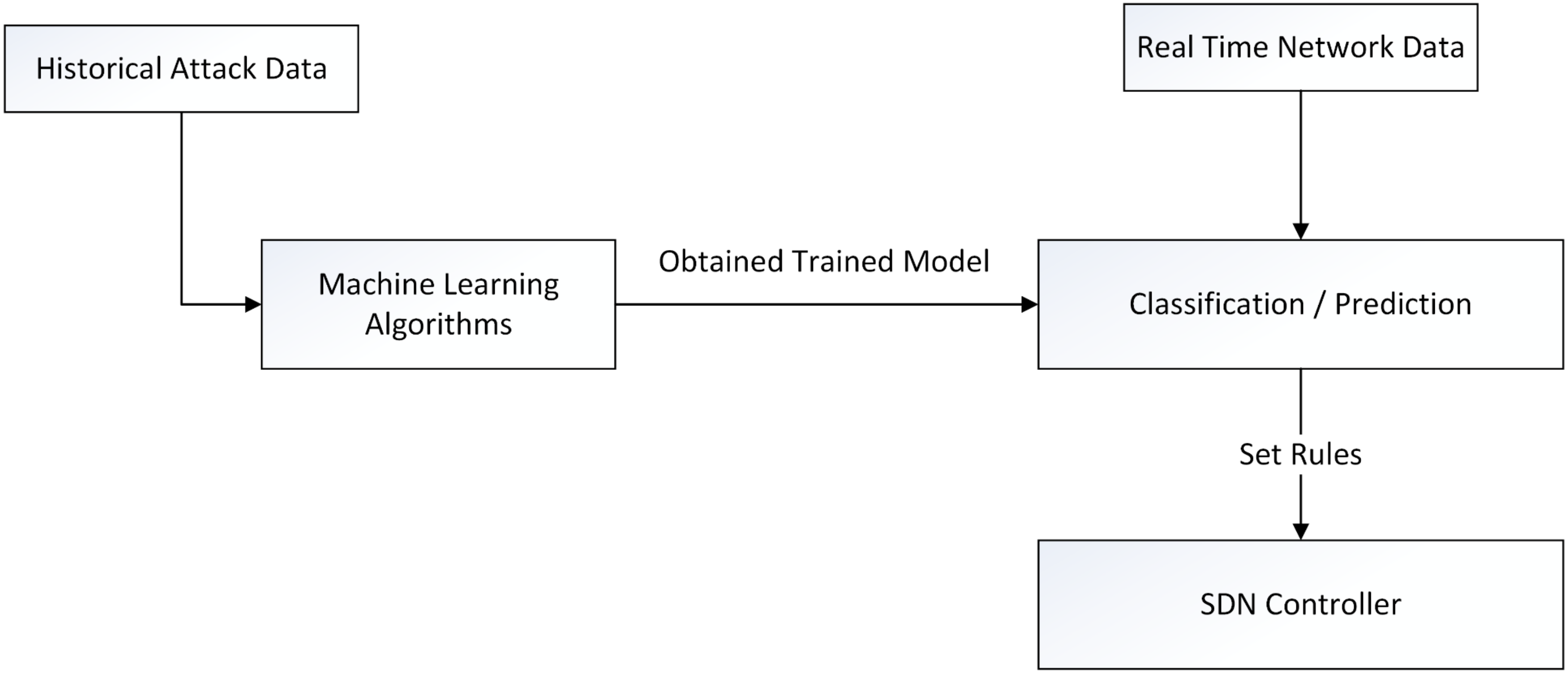
- Abdou et al. recorded ~17M login attempts from 112 different countries and over 6K distinct IP addresses.
- Our Objective
  - To use machine learning algorithms on historical network attack data set to predict the host which will be attacked.
  - To block particular subnet as a whole rather than blocking individual IP addresses.

Methodology

Machine Learning algorithms

- C4.5
  - To generate a decision tree based on information entropy.
- BayesNet
  - Probabilistic graphical model using a directed acyclic graph (DAG) showing a set of random variables and their conditional dependencies.
- Naive-Bayes
  - Probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions between the features.
- Decision Table
  - Compact way to model complex rule sets and respective actions using methods such as: flowcharts, switch-case and if-then-else.

Architecture


Historical Attack Data

Machine Learning Algorithms

Obtained Trained Model

Real Time Network Data

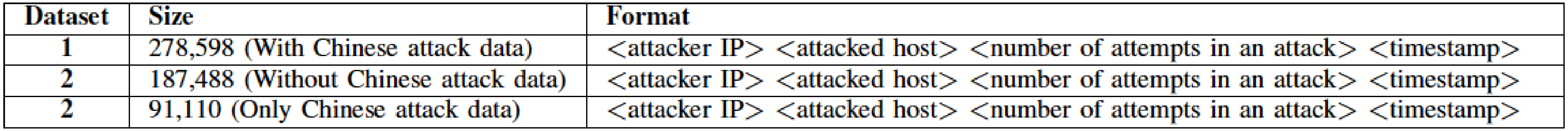
Classification / Prediction

Set Rules

SDN Controller

Dataset

- ❑ **We are using a public dataset from “*LongTail*”**
  - ❑ Open source project that records SSH brute force attacks 32 honeypots.
  - ❑ Dataset 1 - List of attacks including china = 278,598
  - ❑ Dataset 2 - List of attacks without china = 187,488
  - ❑ Dataset 3 – List of attacks with only from china = 91,110



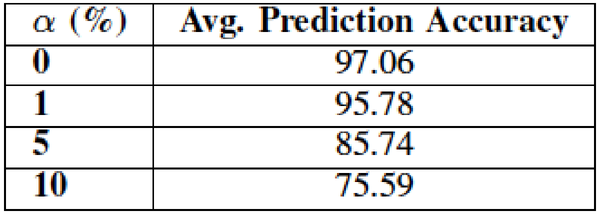
Dataset	Size	Format
1	278,598 (With Chinese attack data)	<attacker IP> <attacked host> <number of attempts in an attack> <timestamp>
2	187,488 (Without Chinese attack data)	<attacker IP> <attacked host> <number of attempts in an attack> <timestamp>
2	91,110 (Only Chinese attack data)	<attacker IP> <attacked host> <number of attempts in an attack> <timestamp>

Experiment Results

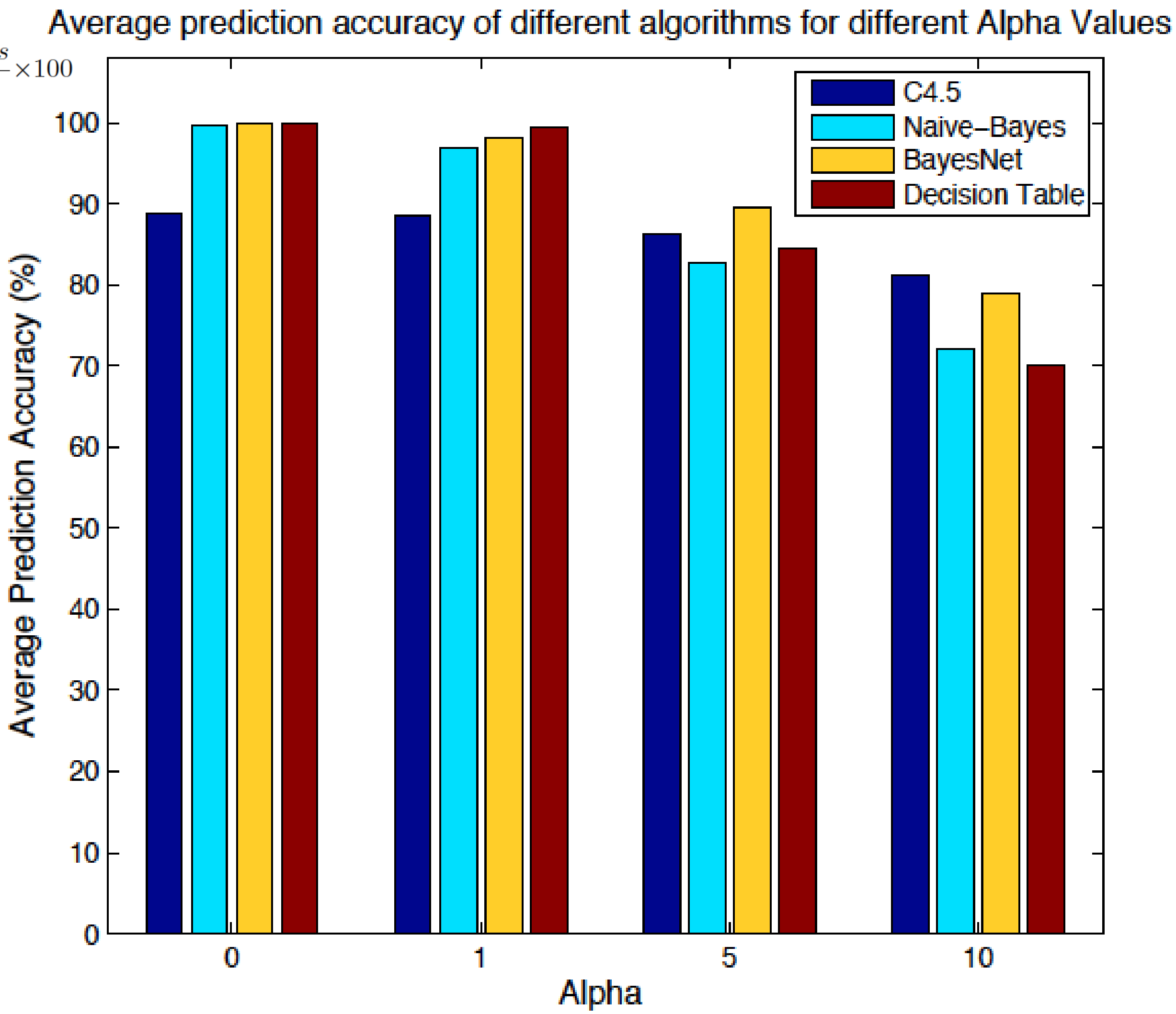
- **Results and discussion.**
  - Weka – Java based ML tool.
  - The datasets were split in 30/70, 40/60, 50/50, 60/40, and 70/30 ratio for training and testing purposes.
  - Prediction accuracy of different ML algorithms, for different data sets, training/testing split ratio and the threshold *α*.
  - *α* is the minimum probability required to consider any host as vulnerable.
  - An average prediction accuracy of 91.68% was achieved with Bayesian Network (254,834 out of 278,598 attacks).
  - Highest accuracy of 99.99% is obtained with Decision Table for dataset 1 when alpha = 0.

Average Prediction Accuracy with Alpha

$$Accuracy = \frac{Number\ of\ correctly\ predicted\ attacks}{Total\ number\ of\ attacks} \times 100$$

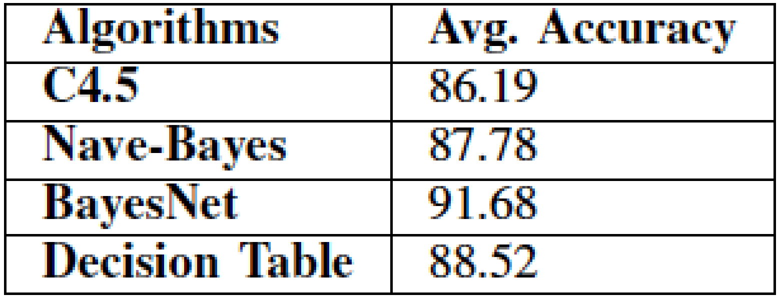


α (%)	Avg. Prediction Accuracy
0	97.06
1	95.78
5	85.74
10	75.59



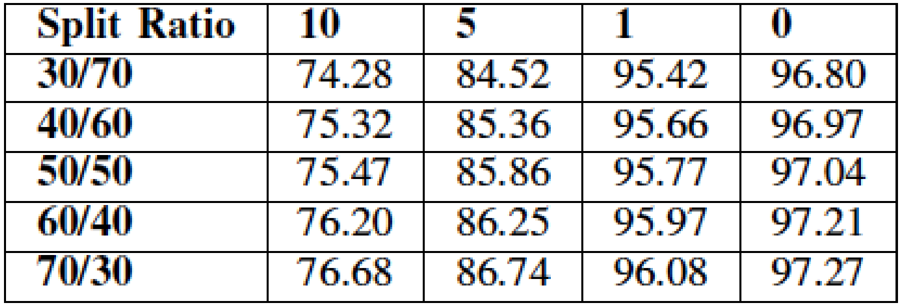
Overall Accuracy and Percentage Split

Average Prediction Accuracy



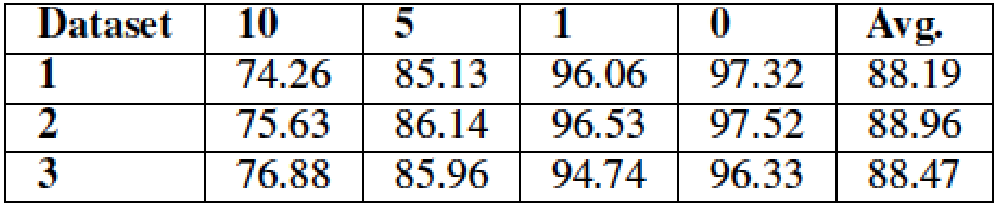
Algorithms	Avg. Accuracy
C4.5	86.19
Nave-Bayes	87.78
BayesNet	91.68
Decision Table	88.52

Prediction Accuracy based on training/testing split

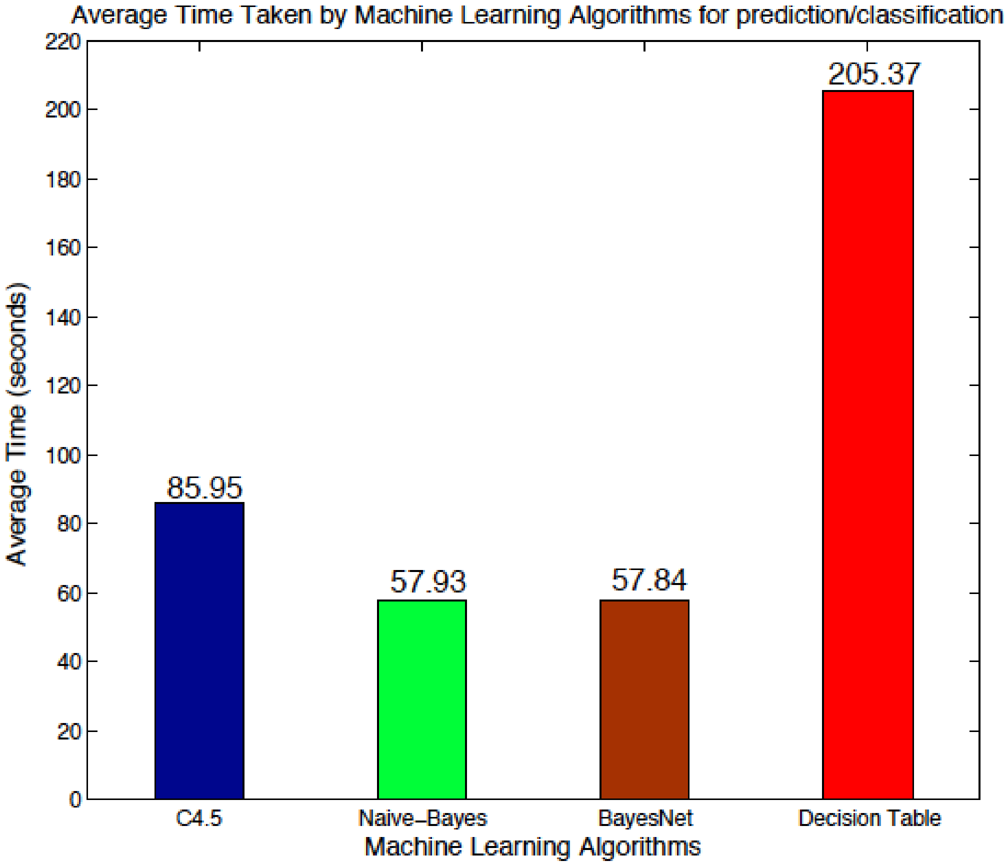


Split Ratio	10	5	1	0
30/70	74.28	84.52	95.42	96.80
40/60	75.32	85.36	95.66	96.97
50/50	75.47	85.86	95.77	97.04
60/40	76.20	86.25	95.97	97.21
70/30	76.68	86.74	96.08	97.27

Effect of Dataset on Average Prediction Accuracy



Dataset	10	5	1	0	Avg.
1	74.26	85.13	96.06	97.32	88.19
2	75.63	86.14	96.53	97.52	88.96
3	76.88	85.96	94.74	96.33	88.47



Conclusion

- Machine Learning approach can help in defining security rules for SDN controller.
- A small probability of attack, obtained through ML approach, has significant effect on the SDN security.
- Achieved an average prediction accuracy of 91.68% with Bayesian Network (total 278,598 attacks).
- Blocking the subnet, rather than the individual IPs.
- The decline in accuracy in response to increasing α proves even small probability of attack cannot be ignored.

GitHub: https://github.com/wedaa/LongTail-Log-Analysis

Thank You!

Questions