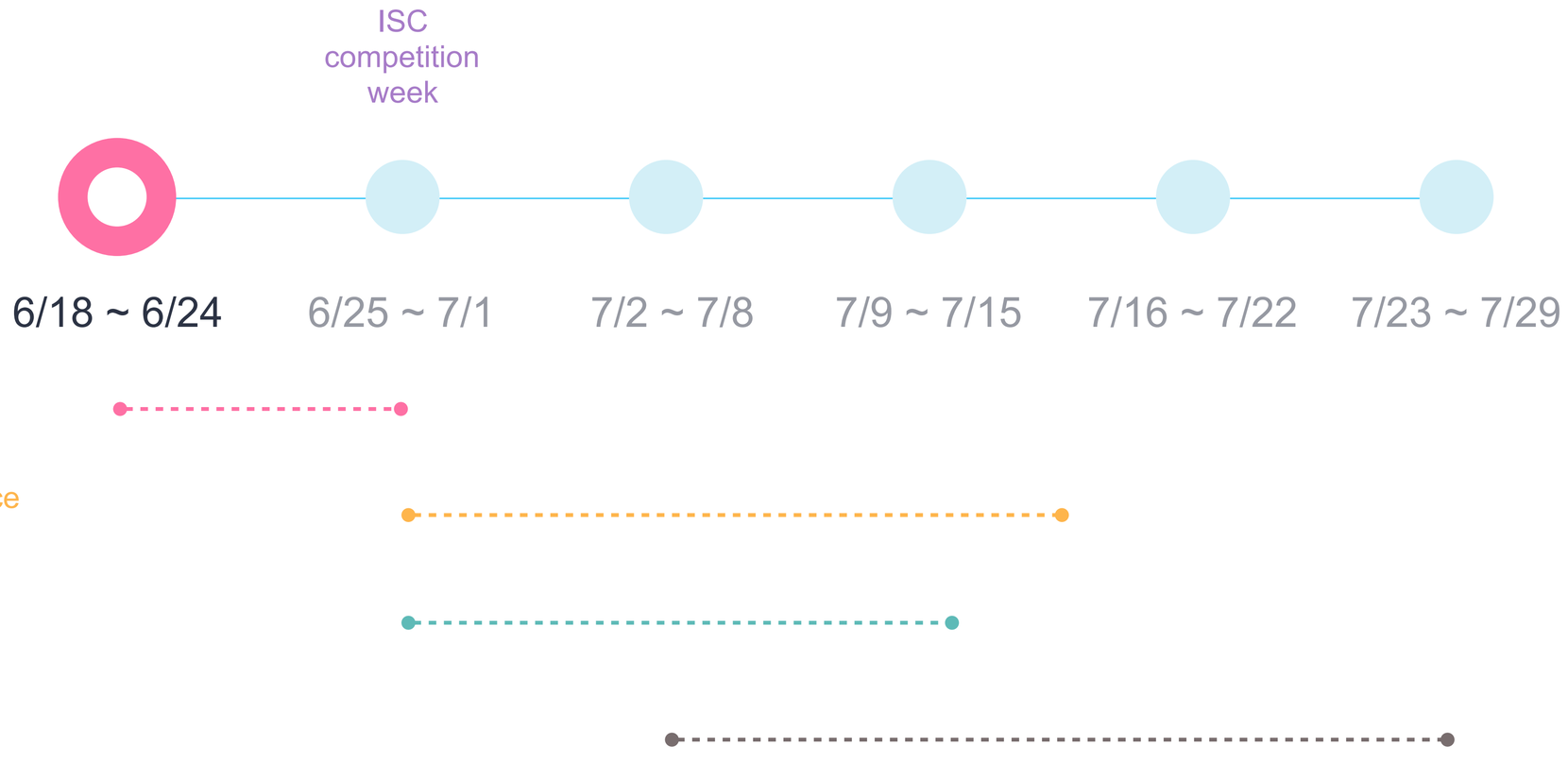
Task weekly **Report**

updated on 20 June 2018

Task plan each **Week**



Planing about **Preprocess Data**

Prepare **Dataset**

Problem

There are many attribute that can’t use directly

for training the model in SVM and Deep learning.

For example,

MAC address 00:01:80:12:15:12

Because SVM and NN not supports purely categorical data. They are supporting data

in vector format and continuous value.

Prepare **Dataset**

Solution

Using Assign Group for Dataset or one-hot technic like,

x = [1, 2, 3, 3, 2, 1]

encoding

1 = 0,0,1

2 = 0,1,0

3 = 1,0,0

Dataset **Grouping**

set 1

Description

Try grouping the attributes that there are categorical data

- IP address
- MAC address
- Port number
- MAC address type
- LLC type
- IP version

Dataset **Grouping**

set 1

How

Assign range of the data

1. IP Address

Private (Group P)

10.0.0.0 – 10.255.255.255
172.16.0.0 – 172.31.255.255
192.168.0.0 – 192.168.255.255

Usage overall

(Group U) | 1.0.0.0 - 126.255.255.255 (A), 128.0.0.0 - 191.255.255.255 (B), 192.0.0.0 - 223.255.255.255 (C)
(Group L) | 127.0.0.0 - 127.255.255.255 (loopback and diagnostic functions)
(Group E) | 224.0.0.0 - 239.255.255.255 *Experiment
(Group M) | 240.0.0.0 - 254.255.255.254 *multicast
(Group BA) | 255.255.255.255 *broadcast all

Ref : http://www.vlsm-calc.net/ipclasses.php, https://www.computerope.com/jargon/i/ip.htm

Dataset **Grouping**

set 1

How

Assign range of the data

2. Port

[pW] Well known port : 0–1023
[pR] Registered ports : 1024–49151
[pP] Private ports : 49152-65535

3. MAC Address

MAC type
[2048] 0x2048 = IPv4
[0] 0x0000 = IEEE802.3
[2054] 0x2054 = ARP
[34525] 0x86DD = IPv6
[24578] 0x24578 = DEC MOP Dump/Load
Ref: https://en.wikipedia.org/wiki/EtherType

MAC range
[mac-type1] : CDP, VTP
[mac-type2] : Cisco
[mac-type3] : IEEE 802.x / Link Layer Discovery Protocol
[mac-type4] : IPv4 multicast
[mac-type5] : IPv6 multicast
[mac-type6] : IEC 61850-8-1 / GSSE / IEC 61850 8-1
[mac-type7] : Local
Ref: https://www.iana.org/assignments/ieee-802-numbers/ieee-802-numbers.xhtm

Dataset **Grouping**

set 1

How

Assign range of the data

4. LLC types

[0] Null LSAP
[66] IEEE 802.1 Bridge Spanning Tree Protocol
[170] SNAP Extension Used

Ref: https://en.wikipedia.org/wiki/IEEE_802.2

Dataset **Grouping**

set 2

Description

Try to convert all of categorical data to be in vector format.

How

Encoding attribute by using library in keras.
Ref: https://machinelearningmastery.com/how-to-one-hot-encode-sequence-data-in-python/

Problem may occurs

MemError in python because of too many attributes.
May be fix by divide training data.

Features **assigned**

Preprocess 1 from dataset get **55** attributes

[**54** attributes **for training**]
1. Ether_or_Dot3 :
    ‘0’ = 802.3/Dot3
    ‘1’ = Ether
[2-8]. smact[1-7] : source MAC address group [1-7]
[9-15]. dmact[1-7] : destination MAC address group [1-7]
16. MAC-2048
17. MAC-0
18. MAC-2048
19. MAC-0
20. MAC-2054
21. MAC-34525
22. MAC-24578
23. LLC :
    ‘0’ = is not llc
    ‘1’ = is llc
24. llc-ssap-0
25. llc-ssap-66
26. llc-ssap-170
27. llc-dsap-0
28. llc-sap-66
29. llc-dsap-170
[30-34],[35-39]. s[P, U, L, E, M, BA], d[P, U, L, E, M, BA]
40. IP_ttl
[41-43]. ip[0,4,6]
44. TCP
[45-47],[48-50]. sp[W, R, P], dp[W, R, P]
51. UDP
52. ARP
53. ICMP
54. pLen
55. Status [**1** attributes **for prediction**]
  ‘0’ = Normal
  ‘1’ = Attack

Experiment **Dataset**

Main point

Find the result that converted data is functional for training and prediction the model.

How

Find the result that dataset is useful. By evaluate the model.

Experiment with **SVM model**

using Sklearn library [try with CPU]

Assign parameters

Input - 54 vectors (attributes)
Output - 1 vector (Status : Attack or not)
c = 0.5 (high avoid misclassifying)
Time used : more than an hour, around 6-8 hours

Dataset used : Training dataset 100,000 packets, Test dataset 100,000 packets

Evaluate technic used : Mean Square Error (MSE) [now.. calculate..]

Using Preprocess Technic : 1 (Grouping port, ip, MAC address …)

Plan to **evaluate**

Description

using confusion matrix to calculate accuracy.

Ref: https://en.wikipedia.org/w…/Evaluation_of_binary_classifiers

Experiment with **Deep learning model**

using Keras library [try with GPU]

Assign parameters

Layers - 3 layers [1] relu [2] sigmoid [3] softmax
Input - 54 vectors (attributes)
Output - 1 vector (Status : Attack or not)
Batch size - 32
Time used - 1~2 minutes

Dataset used : Training dataset 100,000 packets, Test dataset 100,000 packets

Evaluate technic used : Mean Square Error (MSE) = 0.49228

Using Preprocess Technic : 1 (Grouping port, ip, MAC address …)

Summary Tasks on this week

Preprocess Data

- [x] Preprocess set 1
- [ ] Preprocess set 2

Experiment evaluate model

- [x] Testing with Preprocess set 1
- [ ] Testing with Preprocess set 2