한국사물인터넷학회논문지 Vol. 3, No. 1, pp. 9-11, 2017 https://doi.org/10.20465/KIOTS.2017.3.1.009

Hierarchical Clustering을 이용한 네트워크 패킷의 분류

여인성[1], Quan Tran Hai[1], 황성운[2*]
[1]홍익대학교 전자전산공학과, [2]홍익대학교 컴퓨터정보통신공학과

Classification of network packets using hierarchical clustering

Insung Yeo[1], Quan Tran Hai[1], Seong Oun Hwang[2*]
[1]Department of Electronics and Computer Engineering, Graduate School of Hongik University
[2]Department of Computer and Information Communications Engineering, Hongik University

요 약 최근에 인터넷과 모바일 장치가 널리 보급되면서 해커들이 네트워크를 이용해 공격하는 횟수 또한 증가하고 있다. 네트워크를 연결할 때 패킷을 주고받으며 통신을 하게 되는데, 여기에는 다양한 정보가 포함되어 있다. 이 패킷들의 정보를 Hierarchical Clustering 분석을 사용해 분석하고 정상적인 패킷과 비정상적인 패킷을 분류하여 공격자들의 공격을 탐지하였다. 이 분석 방법을 통해 새로운 패킷을 분석하여 공격을 탐지하는 것이 가능할 것이다.

주제어 : 네트워크 보안, clustering

Abstract Recently, with the widespread use of the Internet and mobile devices, the number of attacks by hackers using the network is increasing. When connecting a network, packets are exchanged and communicated, which includes various information. We analyze the information of these packets using hierarchical clustering analysis and classify normal and abnormal packets to detect attacks. With this analysis method, it will be possible to detect attacks by analyzing new packets.

Key Words : Network security; clustering

1. 서론

1.1 문제 상황

최근에 인터넷과 모바일 장치가 널리 보급되면서 이를 이용하는 사용자의 수가 급증하고 있다. 이에 따라 해커들이 네트워크를 이용해 공격하는 횟수 또한 증가하고 있다. 네트워크를 연결할 때 패킷을 주고받으며 통신을 하게 되는데, 여기에는 다양한 정보가 포함되어 있다. 이 패킷들의 정보를 분석하고 정상적인 패킷과 비정상적인 패킷을 분류하여 공격자들의 공격을 탐지하고자 한다.

1.2 자료

사용한 자료는 'KDD Cup 1999 Data'로 데이터 마이닝 대회에서 사용되었다[1]. 이 자료는 군용 네트워크 환경에서 시뮬레이션한 다양한 종류의 패킷정보를 포함하고 있다. 데이터셋을 그대로 사용하기에 크기가 커서 정상패킷 5618개, 공격패킷 3827개를 임의로 추출하였다.

2. 분석 기법

2.1 패킷에 시그널 변수 추가

*교신저자 : 황성운(sohwang@hongik.ac.kr)
접수일 2017년 2월 28일 심사완료일 2017년 3월 21일

자료를 정상패킷(normal)과 공격패킷(abnormal)으로 구분하고 각각에 시그널 변수를 붙인다. 시그널이 1이면 정상패킷이고, 시그널이 0이면 공격패킷이다. 다음은 R을 이용하여 시그널 변수를 추가하는 과정을 나타낸 것이다.

```
# 데이터셋 삽입
my<-read.csv("C:/TCP_u2r_r2l.csv")

# 정상과 공격패킷을 분리
my.n<-subset(my,attack_type=="normal.")
table(my.n$attack_type)
my.a<-subset(my,attack_type!="normal.")
table(my.a$attack_type)

# 정상과 공격패킷에 시그널을 추가
n.s<-rep(1,nrow(my.n)) #n.s=1 if it is normal
my.n<-data.frame(my.n,n.s)
n.s<-rep(0,nrow(my.a)) #n.s=0 if it is abnormal
my.a<-data.frame(my.a,n.s)

# 정상, 공격 패킷을 병합
my<- rbind(my.n, my.a)
```

[Fig. 1] Adding a Signal Variable

2.2 변수 선택

다중선형회귀 분석과 단계적선형회귀(stepwise) 분석으로 주요변수를 추출한다. α(유의수준)을 0.05로 정하고 P-value가 0.05 이상인 변수를 삭제한다.

```
duration                -2.017e-06  2.756e-07  -7.318 2.72e-13 ***
src_bytes               -8.638e-10  1.041e-10  -8.299  < 2e-16 ***
dst_bytes                1.078e-08  4.173e-09   2.582 0.009835 **
hot                     -1.631e-02  4.485e-04 -36.360  < 2e-16 ***
num_failed_logins       -7.857e-02  1.193e-02  -6.585 4.79e-11 ***
logged_in                9.030e-02  1.377e-02   6.555 5.84e-11 ***
num_compromised         -1.385e-02  4.876e-03  -2.840 0.004525 **
root_shell              -2.866e-01  1.339e-02 -21.404  < 2e-16 ***
su_attempted            -1.923e-01  6.398e-02  -3.006 0.002651 **
num_root                 1.229e-02  3.232e-03   3.803 0.000144 ***
num_file_creations      -1.649e-02  2.774e-03  -5.943 2.89e-09 ***
is_guest_login           8.766e-02  2.491e-02   3.519 0.000435 ***
count                    4.956e-05  1.943e-05   2.550 0.010785 *
srv_count                1.269e-04  8.021e-05   1.582 0.113752
rerror_rate             -1.306e-01  1.690e-02  -7.727 1.22e-14 ***
```

[Fig. 2] Output of StepAIC

2.3 클러스터링으로 분류

Hierarchical Clustering 분석을 사용해 데이터셋을 계층관계와 상호 유사성에 기초하여 군집화 한다. 우선 클러스터 수(k)를 100개로 정하여 실행하였고 각 Cluster Identity에 속하는 정상/공격패킷의 수를 출력하였다(그림4).

```
# Hierarchical Clustering
hc <- hclust(dist(my[,3:27]), "ave")
c.n <- cutree(hc, k=100)
my2<-data.frame(my, c.n)

# 정상패킷과 공격패킷의 Cluster Identity 출력
table(my2$c.n, my2$n.s==1)
```

[Fig. 3] Clustering

Cluster	normal	abnormal
1	33	87
2	0	6
3	0	59
4	0	1782
5	0	1014
6	0	1202
7	0	862
8	0	1
9	1	1
10	29	9
11	189	45
12	2	12
13	0	8
14	0	3
15	0	5
16	0	1
17	0	13
18	0	3
19	523	34
20	2	10

Cluster	normal	abnormal
21	0	11
22	0	199
23	5	1
24	10	10
25	0	1
26	0	3
27	0	2
28	0	1
29	0	4
30	0	7
31	0	2
32	0	7
33	0	8
34	0	23
35	0	20
36	0	4
37	0	9
38	0	3
39	0	10
40	0	9

Cluster	normal	abnormal
41	0	11
42	0	8
43	0	2
44	0	2
45	7	1
46	7	3
47	1	1
48	59	1
49	0	2
50	0	5
51	0	1
52	0	4
53	24	1
54	0	2
55	0	3
56	0	5
57	0	13
58	0	1
59	0	1
60	0	15

Cluster	normal	abnormal
61	1	1
62	0	1
63	0	5
64	0	9
65	1	2
66	0	14
67	0	1
68	0	2
69	0	1
70	0	3
71	0	2
72	0	1
73	0	1
74	0	2
75	0	1
76	0	1
77	0	1
78	2516	0
79	1	0
80	1	0

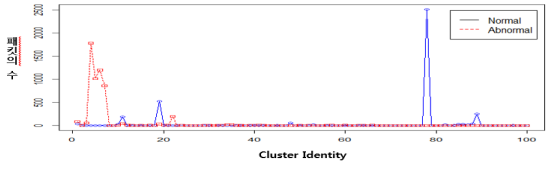
Cluster	normal	abnormal
81	1	0
82	22	0
83	3	0
84	7	0
85	29	0
86	28	0
87	19	0
88	39	0
89	251	0
90	1	0
91	1	0
92	3	0
93	3	0
94	3	0
95	2	0
96	1	0
97	6	0
98	1	0
99	1	0
100	1	0

[Fig. 4] Output of hclust

3. 결과

3.1 정상패킷과 공격패킷의 Cluster Identity 그래프

클러스터 수가 100일 때 각각의 Cluster가 정상패킷인지 공격패킷인지를 판별하기 위해 그림5의 Cluster Identity 그래프를 그려보았다. 정상패킷은 Cluster 78에 대부분 분포되어 있고 공격패킷은 Cluster 4, 5, 6, 7에 많이 분포되어 있음 알 수 있다.



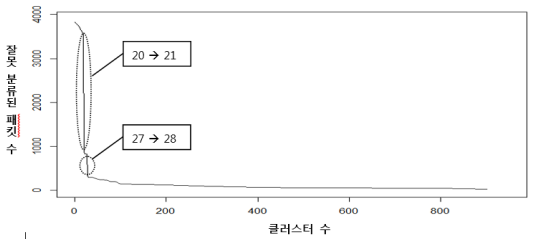
[Fig. 5] Graph of Cluster Identity

3.2 최적 클러스터 수 결정

클러스터 수에 따라서 잘못 분류된 패킷의 수를 측정하기 위해 전체 패킷 수 9445개의 분류 정확도(%)를 측정하였다(Table1). 클러스터의 수가 20에서 21로 증가할 때, 27에서 28로 증가할 때 잘못 분류된 패킷의 수가 현저하게 줄어들며 이에 따라 분류 정확도가 급격히 상승하는 것을 확인할 수 있다(그림6).

〈Table 1〉 Classification accuracy of each cluster

클러스터 수	20	21	...	27	28
오분류 패킷	3553	886	...	824	339
분류 정확도	62.38%	90.61%	...	91.27%	96.41%



[Fig. 6] Graph of Misclassified packets

4. 결론

위의 분석 결과를 통해, 클러스터링 분석을 이용하여 네트워크 패킷을 정상, 비정상으로 분류하는 방법을 알아보았고 특정 클러스터의 수에서 분류정확도가 급격히 상승하는 것을 확인하였다. 이 분석방법을 좀 더 보완한다면 새로운 패킷을 분석하여 정상, 비정상으로 판단하는 것이 가능할 것이다. 다만, 정상패킷과 유사하여 구분이 되지 않는 공격패킷(예: SYN flood)은 이 방법으로 정상 여부를 판단하기 어려울 것으로 보인다.

ACKNOWLEDGMENTS

이 논문은 2017년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구이다(No. 2017R1A2B4001801).

REFERENCES

[1] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
[2] http://cran.r-project.org/manuals.html
[3] http://blogchannel.tistory.com/44
[4] https://rpubs.com/cardiomoon/18980

여 인 성(Insung Yeo) [비회원]

- 2015년 5월 : 홍익대학교 컴퓨터정보통신학과 (학사)
- 2017년 2월 : 홍익대학교 전자전산공학과 (석사)

<관심분야>
사이버보안

Quan Tran Hai(Quan Tran Hai) [비회원]

- 2011년 10월 : 베트남 University of science 전자통신학과 (학사)
- 2017년 2월 : 홍익대학교 전자전산공학과 (석사)

<관심분야>
사이버보안

황 성 운(Seong Oun Hwang) [비회원]

- 1998년 2월 : 포항공과대학교 정보통신학과 (석사)
- 2004년 8월 : 한국과학기술원 전자전산학과 (박사)
- 2006년 1월 ~ 2006년 12월 : University of Michigan 박사후 연구원
- 2008년 3월 ~ 현재 : 홍익대학교 컴퓨터정보통신학과 부교수

<관심분야>
정보보호, 암호