패턴인식 기반 침입탐지를 위한 데이터셋 구성 기법에 대한 연구*

공성현*, 조민정*, 조재익**, 이창훈*†
*서울과학기술대학교 컴퓨터공학과
**IBM
e-mail : gongsh@seoultech.ac.kr

A Study on Dataset Construction Technique for Intrusion Detection based on Pattern Recognition

Seong-Hyeon Gong*, Min-Jeong Cho*, Jae-ik Cho**,Changhoon Lee* †
*Dept. of Computer Engineering, Seoul National University of Science and Technology
**IBM

요 약

통신 기술이 발달하고, 네트워크 환경 또한 다양해짐에 따라 통신 사용자들에 대한 사이버 위협 또한 다양해졌다. 패턴인식 기술과 기계학습에 기반한 침입탐지 기술은 새롭게 보고되는 수많은 사이버 공격들에 대응하기 위해 등장하였다. 기계학습 기반의 IDS 는 낮은 오탐률과 높은 효율성을 요구하며, 이러한 특징은 데이터셋을 구성하는 방법론에 큰 영향을 받는다. 본 논문에서는 패턴인식 기반 트래픽 분석을 수행하기 위한 데이터셋을 구성할 때 고려해야 할 주안점에 대해 논하며, 현실의 사이버 위협 상황을 잘 반영할 수 있는 데이터셋을 도출하는 방안을 모색한다.

1. 서론

네트워크를 구성하던 단말 장치의 수가 통신 기술의 발달로 급격히 증가함에서 따라 네트워크 프로토콜 및 사이버 위협의 종류 또한 다양해졌다. 이러한 변화는 사이버 침해 탐지 기술을 기존의 시그니처 기반의 패턴인식 기반 시스템에서 기계학습에 기반한 지능형 시스템으로 변화시켰다. 네트워크 트래픽 분석은 사이버 공격 탐지를 위한 대표적인 방법론으로, 적절한 데이터셋과 적합한 알고리즘을 적용할 경우, 높은 정탐률과 효율성을 기대할 수 있는 탐지 기법이다. 그러나, 적절한 데이터셋과 적합한 알고리즘을 선택하는 과정은 정확한 이론/실험적 근거를 필요로 한다. 알고리즘을 선택하는 과정은 주어진 데이터를 최대한 활용할 수 있도록 많은 실험적 결과를 요구하며, 데이터셋의 경우, 데이터셋의 차원과 각 데이터들의 특성은 향후 적용될 알고리즘의 정확성과 효율에 큰 영향을 줄 수 있기 때문에 이들을 적절히 구성하고 선택하기 위한 많은 연구가 필요하다.

본 논문은 알고리즘의 정확도와 효율성을 높이고 현실의 상황을 잘 반영할 수 있는 데이터셋을 구성하기 위해서 논의되어야 할 주안점들에 대하여 분석하고, 적절한 데이터셋을 구성하기 위한 방안에 대하여 살펴본다. 2 장에서는 과거의 데이터셋 구성 사례를 통해 데이터셋 구성시 주의해야할 문제점들에 대하여 살펴보고, 3 장에서는 과거에 소개된 문제점들을 해결하기 위한 방안들에 대하여 분석한다. 4 장에서는 적절한 데이터셋 도출을 위한 신규 방안을 제안하며, 5 장에서 결론을 맺는다.

2. 네트워크 트래픽 분석용 데이터셋 구성 사례

KDD(Knowledge Discovery in Database) CUP 1999 DataSet 은 데이터마이닝 및 지식발견 대회에서 사용되었으며 컴퓨터 네트워크의 침입 탐지를 주제로 하는 데이터셋이다[1]. MIT 의 Lincoln Lab 에 의해 생성된 이 데이터셋은 미 공군의 가상 내부망 환경에서 9 주동안 정상 접근과 공격 상황을 시뮬레이션한 패킷 데이터들을 수집한 후 41 개의 성분을 추출하여 재구성한 결과물이다. 이 데이터셋은 24 개의 공격 유형들을 포함한 트레이닝셋과 14 개의 신규 공격을 포함한 테스트셋을 가지고 있으며, TCP 커넥션단위로 재구성된, 약 700 만개 데이터를 포함하고 있다. 각 공격유형들은 DoS(Denial of Service), R2U(Root to User), U2R(User to Root), Probing 등 4 개중 하나의 공격 유형으로 분류된다. 다음 표는 각 공격유형들의 세부 공격 기법 및 공격 빈도수를 나타낸 표이다.

*: 본 연구는 2016 년도 산업통상자원부의 재원으로 한국 에너지기술평가원 (KETEP)의 지원을 받아 수행한 연구과제입니다. (No. 20161510101810)
†: 교신저자(chlee@seoultech.ac.kr)

<표 1> KDD’99 공격 종류 및 빈도 수

Attack Type	Attacks	#num of attacks
DoS	smurf, neptune, back, teardrop, pod, land	391,458
R2L	warezclient, guess_passwd, warezmaster, imap, ftp_write, multihop, phf, spy	1,126
U2R	buffer_overflow, rootkit, loadmodule, perl	52
Probing	satan, ipsweep, portsweep, nmap	4,107

3. 데이터셋 구성의 주안점

3.1 데이터셋의 크기

데이터 마이닝을 통한 분석 정확도를 최대한으로 유지하면서 동시에 데이터셋의 크기를 최소화하여 분석에 필요한 연산의 수를 줄이는 것은 데이터셋 구성에 있어 가장 중요한 주안점중 하나이다. 이를 위해 등장한 다양한 Feature Selection 기법들은 전체 데이터를 가장 잘 대표할 수 있는 feature 들로만 데이터셋을 구성하여 데이터셋의 차원을 축소시킨다. 실제로 KDD’99 데이터셋의 경우, 41 개의 feature 들로 구성되어있지만 공격 탐지의 대상이 되는 4 개의 공격 종류들 중 하나로 대상을 좁힐 경우, 분석에 큰 영향을 주는 주요 성분은 크게 줄어들게 된다. <표 2>는 정상 트래픽과 공격 트래픽 사이의 관계와, 정상트래픽과 특정 분류의 공격 트래픽 사이의 유효한 관계 유무를 분석한 자료이다.

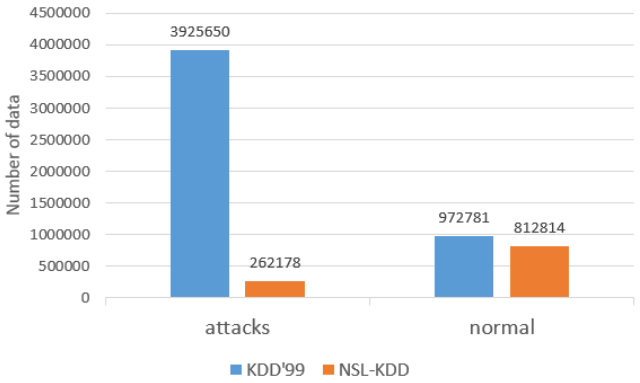
<표 2> KDD’99 의 공격 종류와 특징간 연관관계 [4]

No	Name	Attack	DoS	Probe	R2L	U2R
1	duration				O	
2	protocol_type					
3	service	O			O	O
4	flag			O		
5	src_bytes	O	O	O	O	O
6	dst_bytes	O	O		O	O
7	land		O			
8	wrong_fragment		O			
9	urgent					
10	hot				O	
11	num_failed_logins				O	
12	logged_in	O				
13	num_compromised					
14	root_shell					O
15	su_attempted					
16	num_root					O
17	num_file_creations					
18	num_shells					
19	num_access_files					
20	outbound_cmds					
21	is_hot_login					O
22	is_guest_login				O	
23	count	O			O	
24	srv_count	O				O
25	serror_rate	O				
26	srv_serror_rate	O				
27	rerror_rate			O		
28	srv_error_rate			O		
29	same_srv_rate					
30	diff_srv_rate		O	O		
31	srv_diff_host_rate	O				
32	dst_host_count	O				
33	dst_host_srv_count	O				
34	dst_host_same_srv_rate					
35	dst_host_diff_srv_rate	O				
36	dst_host_same_src_port_rate		O	O		O
37	dst_host_srv_diff_host_rate			O		
38	dst_host_serror_rate					
39	dst_host_srv_serror_rate		O		O	
40	dst_host_error_rate			O		
41	dst_host_srv_rerror_rate			O		

<표 2>를 바탕으로 특정 공격 기법에 초점을 맞춰 분석을 수행할 경우, 데이터셋의 특징 수를 7~9 개 정도로 축소할 수 있으며, 역으로 공격 탐지와 관련성이 낮은 특징을 추출하여 전체 데이터셋의 차원을 축소할 수도 있다[4].

3.2 중복 데이터로 인한 편향

KDD’99 데이터셋의 분포를 보면, 4,898,431 개의 트레이닝 데이터셋중 공격으로 분류된 데이터의 비중이 93%에 이른다. 테스트셋에서도 역시 88%라는 높은 수치를 보이는데, 이는 일반적으로 예상할 수 있는 네트워크 환경에 비해 공격의 빈도가 지나치게 높은 것을 확인할 수 있다. 이러한 데이터들 중 상당수의 데이터들은 중복된 데이터인데, 기계학습 알고리즘들은 많은 빈도수를 보이는 패턴으로 주어진 데이터를 판별하려는 편향 특성을 보이기 때문에 중복되는 데이터들은 트래픽 분석의 정확도에 악영향을 미친다[2]. 또한, 분석의 결과가 빈도수에 편향되는 특성은 공격 종류를 구분하는 분석의 정확도에도 영향을 미치는데, 상대적으로 많은 패킷을 요구하는 DoS 공격은 탐지가 쉬운 반면 적은 패킷을 요구하지만, 공격의 위험성을 훨씬 높은 U2R, R2L 공격의 경우, 공격에 사용된 트래픽의 수가 다른 공격들에 비해 적기 때문에 이들에 대한 탐지율이 낮아지는 현상이 발생한다[3]. 이러한 문제를 해결하기 위해 KDD’99 데이터셋의 개선된 버전인 NSL-KDD 데이터셋은 상대적으로 빈도수가 높던 공격 데이터의 수를 크게 축소하고, 빈도수가 낮은 노멀 데이터의 수를 적게 축소하여 일반적인 네트워크 환경에서의 트래픽 분포와 유사하도록 데이터셋을 새롭게 구성하였다[3].



(그림 1) KDD’99 와 NSL-KDD 의 데이터 분포

3.3 서브셋 구성에 따른 분석결과의 차이

크기가 큰 데이터셋을 가지고 마이닝을 수행할 경우 분석가는 전체 데이터의 일부를 서브셋 형태로 추출하여 분석에 사용할 수 있다. KDD’99 의 원본 트레이닝셋은 5 백만개 정도의 데이터를 보유하고 있지만 이들을 모두 트레이닝으로 사용하기엔 연산의 양이

적지 않아 일반적으로 10%정도의 서브셋을 추출하여 트레이닝에 사용한다. 이 때, 서브셋을 전체 데이터셋에서 랜덤하게 추출할 경우 서브셋 내의 데이터가 어떻게 추출되는가에 따라 분석 결과가 크게 달라질 수 있다. Gharibian 과 Ghorbani[5]의 연구에 따르면, 서브셋을 구성하는 노멀 데이터와 공격 데이터의 비율을 다양하게 구성할 경우, Gaussian 이나 Maive Bayes 등의 확률적 기법들은 탐지율에 대하여 적은 편차를 보이지만, Decision Tree 나 Random Forest 는 비교적 큰 편차를 보인다[5, 6].

<표 3> 서브셋 구성에 따른 알고리즘별 탐지율 [5]

Algorithm	Subset		DoS	Prob	R2L	U2R
	Normal	Attack				
Gaussian	80%	20%	0.967	0.866	0.135	0.461
	88%	12%	0.828	0.844	0.135	0.457
Naive Bayes	80%	20%	0.791	0.805	0.097	0.766
	88%	12%	0.791	0.866	0.097	0.761
Decision Tree	80%	20%	0.887	0.681	0.054	0.171
	88%	12%	0,509	0.678	0.05	0.17
Random Forest	80%	20%	0.827	0.718	0.033	0.183
	88%	12%	0.512	0.712	0.037	0,174

3.4 시계열 기반 트래픽 특성의 영향

트래픽 정보는 시간에 따라 발생하는 정보이기 때문에 시간적 연속성을 고려하여 데이터셋을 구성하여야 한다. 트래픽의 연속성은 트래픽을 전달받는 대상에 대한 관점과 트래픽이 수행하는 서비스에 대한 관점으로 구분할 수 있는데, 전자는 동일 호스트 특징, 후자는 동일 서비스 특징으로 정의된다[2, 7]. 정적 혹은 동적 윈도우 사이즈에 의해 도출된 데이터셋의 각 데이터들은 이러한 시계열 특성을 고려하지 않을 경우, slow-probing 공격과 같이 오랜 시간에 걸쳐 천천히 수행되는 공격을 탐지하지 못할 가능성이 높다. 하나의 트래픽 데이터를 결정짓는 윈도우는 대게 몇 초 이내로 제한되는데, 트래픽을 이용한 공격의 간격이 윈도우 사이즈보다 길 경우 공격의 패턴을 찾지 못하는 문제점을 해결하기 위해 데이터셋은 동일 호스트에 대한 트래픽 정보와 동일 서비스에 대한 트래픽 정보를 충분히 포함함으로써 희소한 공격에 대한 탐지율을 확보할 수 있어야 한다.

3.5 트래픽 충돌에 의한 영향

빠른 속도로 발전하고 있는 모바일 디바이스들로 인하여 무선네트워크 환경의 트래픽 대역폭은 매년 큰 폭으로 증가하고 있다. 이러한 변화는 네트워크상의 트래픽 충돌 확률을 높이는데, 트래픽 충돌은 트래픽의 지연을 유도하여 실제 상황과 다른 트래픽 데이터가 수집되는 상황을 야기한다. 고정된 시간 길이를 갖는 윈도우를 기반으로 수집된 트래픽의 시계열 특성 데이터는 이러한 트래픽 지연에 영향을 받게 되는데, 동일한 윈도우에 포함되어야 하는 트래픽들이 지연으로 인해 서로 다른 윈도우에 포함되게 하여 비정상적인 데이터 및 분석결과를 만들어낼 수 있다.

$$p = 1 - (1 - 1 / W_{avg})^{n-1} \quad \text{(식 1)}$$

네트워크 환경에서 트래픽 충돌이 발생하는 확률은 경쟁 윈도우의 크기에 반비례하며 네트워크 노드 수에 비례하기 때문에[8], 트래픽 데이터 수집을 위한 윈도우의 사이즈를 결정할 때 적합한 윈도우 사이즈가 결정될 수 있도록 네트워크의 환경을 고려해야 한다. (식 1)은 트래픽 충돌 발생 확률에 대한 수식으로, W_{avg} 는 평균 경쟁 윈도우 크기를, n 은 네트워크 노드의 수를 나타낸다.

4. 결론

본 논문에서는 정확하고 효율적인 침입 탐지 분석을 수행하기 위한 데이터셋을 구성할 때 고려해야 할 주안점들을 분석하였다. 패턴인식을 이용한 트래픽 분석은 데이터셋의 크기와 편향성, 시계열 특성을 고려하여 수행되어야 하며 데이터셋의 성향에 맞으면서 동시에 최적의 결과를 도출해낼 수 있는 알고리즘을 선정하기 위해 다양한 알고리즘들을 이용한 분석이 수행되어야 한다. 향후 연구에서는 위에서 언급된 주안점들이 실제 분석 환경에서 여러 패턴인식 알고리즘들에게 어느 정도의 영향을 미치는지에 대한 실험을 통하여 효율적인 데이터셋 특징 추출 기법, 데이터셋의 편향성을 제거하기 위한 방법론 및 시계열 특성을 충분히 반영할 수 있는 최적 윈도우 사이즈 등 데이터셋 구성 방안에 대한 새로운 실험결과를 도출하고자 한다.

참고문헌

[1] KDD CUP 1999 Data, The UCI KDD Archive, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[2] Sahu, Santosh Kumar, Sauravranjan Sarangi, and Sanjaya Kumar Jena. "A detail analysis on intrusion detection datasets." Advance Computing Conference (IACC), 2014 IEEE International. IEEE, 2014.

[3] NSL-KDD dataset, University of New Brunswick, http://www.unb.ca/cic/research/datasets/nsl.html

[4] Zargari, Shahrzad, and Dave Voorhis. "Feature Selection in the Corrected KDD-dataset." Emerging Intelligent Data and Web Technologies (EIDWT), 2012 Third International Conference on. IEEE, 2012.

[5] Gharibian, Farnaz, and Ali A. Ghorbani. "Comparative study of supervised machine learning techniques for intrusion detection." Communication Networks and Services Research, 2007. CNSR'07. Fifth Annual Conference on. IEEE, 2007.

[6] Arora, I. Singh, G. K. Bhatia, and A. P. Singh. "Comparative Analysis of Classification Algorithms on KDD'99 Data Set." International Journal of Computer Network and Information Security (IJCNIS) 8.9 (2016): 34.

[7] Buczak, Anna L., and Erhan Guven. "A survey of data mining and machine learning methods for cyber security intrusion detection." IEEE Communications Surveys & Tutorials 18.2 (2016): 1153-1176.

[8] Bulhões, Rodolfo P., D. Passos, and C. VN Albuquerque. "Collision probability estimation in wireless networks." Communications (LATINCOM), 2016 8th IEEE Latin-American Conference on. IEEE, 2016.