Journal of the Korea Convergence Society
Vol. 10. No. 5, pp. 35-42, 2019

ISSN 2233-4890
https://doi.org/10.15207/JKCS.2019.10.5.035

특징선택 기법에 기반한 UNSW-NB15 데이터셋의 분류 성능 개선

이대범[1], 서재현[2*]
[1]목원대학교 조교수, [2]원광대학교 컴퓨터·소프트웨어공학과 조교수

Classification Performance Improvement of UNSW-NB15 Dataset Based on Feature Selection

Dae-Bum Lee[1], Jae-Hyun Seo[2*]
[1]Assistant Professor, Mokwon University
[2]Assistant Professor, Division of Computer Science & Engineering, Wonkwang University

요 약 최근 사물인터넷과 다양한 웨어러블 기기들이 등장하면서 인터넷 기술은 보다 편리하게 정보를 얻고 업무를 수행하는데 기여하고 있으나 인터넷이 다양한 부분에 이용되면서 공격에 노출되는 Attack Surface 지점이 증가하고 있으며 개인정보 획득, 위조, 사이버 테러 등 부당한 이익을 취하기 위한 목적의 네트워크 침입 시도 또한 증가하고 있다. 본 논문에서는 네트워크에서 발생하는 트래픽에서 비정상적인 행동을 분류하기 위한 희소클래스의 분류 성능을 개선하는 특징선택을 제안한다. UNSW-NB15 데이터셋은 다른 클래스에 비해 상대적으로 적은 인스턴스를 가지는 희소클래스 불균형 문제가 발생하며 이를 제거하기 위해 언더샘플링 방법을 사용한다. 학습 알고리즘으로 SVM, k-NN 및 decision tree를 사용하고 훈련과 검증을 통하여 탐지 정확도와 RMSE가 우수한 조합의 서브셋들을 추출한다. 서브셋들은 래퍼 기반의 실험을 통해 재현률 98%이상의 유효성을 입증하였으며 DT_PSO 방법이 가장 우수한 성능을 보였다.

주제어 : 침입탐지, 특징선택, 데이터 전처리, 희소 클래스, 기계학습

Abstract Recently, as the Internet and various wearable devices have appeared, Internet technology has contributed to obtaining more convenient information and doing business. However, as the internet is used in various parts, the attack surface points that are exposed to attacks are increasing, Attempts to invade networks aimed at taking unfair advantage, such as cyber terrorism, are also increasing. In this paper, we propose a feature selection method to improve the classification performance of the class to classify the abnormal behavior in the network traffic. The UNSW-NB15 dataset has a rare class imbalance problem with relatively few instances compared to other classes, and an undersampling method is used to eliminate it. We use the SVM, k-NN, and decision tree algorithms and extract a subset of combinations with superior detection accuracy and RMSE through training and verification. The subset has recall values of more than 98% through the wrapper based experiments and the DT_PSO showed the best performance.

Key Words : IDS, Feature selection, Data preprocessing, Rare class, Machine learning

*This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP; Ministry of Science, ICT & Future Planning) (No. NRF-2017R1C1B5018128)
*Corresponding Author : Jae-Hyun Seo(delphi7@wku.ac.kr)
Received April 5, 2019 Revised April 28, 2019
Accepted May 20, 2019 Published May 28, 2019

1. 서론

최근 사물인터넷과 다양한 웨어러블 기기들이 등장하면 서 인터넷 기술은 보다 편리하게 정보를 얻고 업무를 수행하는데 기여하고 있다. 하지만 인터넷이 다양한 부분에 이용되면서 공격에 노출되는 공격지점(Attack Surface)이 증가하고 있으며 개인정보 획득, 위조, 사이버 테러 등 부당한 이익을 취하기 위한 목적의 네트워크 침입 시도 또한 증가하고 있다. 시간이 지남에 따라 공격수법이 지능화되는 추세를 보이고 있어 이에 대응하기 위한 많은 연구자들이 다양한 대응 방법을 연구하고 있다. 이러한 인터넷 서비스에 대한 지능화되고 있는 침입을 대응하기 위한 일반적인 방법에는 방화벽과 침입탐지시스템 등이 대표적인 대응방법으로 알려져 있다. 방화벽은 침입으로 의심되는 패킷을 차단하는 역할을 하지만 모든 침입을 차단할 수 없기 때문에 침입탐지시스템에서의 역할이 많은 요구사항을 위해 증가하고 있으며 공격을 식별하고 네트워크 관리자에게 경고함으로써 네트워크 활동을 모니터하고 필터링하는 침입 탐지 시스템 (IDS)이 개발되었다. 데이터 마이닝, 기계 학습, 통계 분석, 유전자 알고리즘, 인공 신경망, 퍼지 논리, 군단 지능 등과 같은 인공지능 기법을 사용하여 다양한 IDS 접근 방식이 등장하였다. 본 논문에서는 KDD'99 데이터셋을 계량한 UNSW-NB15 데이터셋을 이용하였고 네트워크에서 발생하는 트래픽에서 비정상적인 행동을 분류하기 위한 최적의 탐지모델을 제안하고자 머신러닝 기술의 SVM, *k*-NN 및 Decision Tree를 사용하였다. 데이터 전처리를 위해 정상과 비정상 데이터의 클래스 불균형을 조정하였고 위에서 언급한 3가지 특징선택 알고리즘을 적용하여 탐지 정확도와 오탐지율, 오탐률을 평가함으로써 ROC curve 의 탐지정확도에서 98% 이상의 유효성을 입증하였다.

2장에서 인공지능 기반의 침입탐지와 관련된 기존 연구들을 알아보고, 3장에서는 UNSW-NB15 데이터셋에 대하여 살펴본다. 4장에서는 데이터 전처리 과정 및 사용 알고리즘에 대해 알아보고, 5장에서 실험 및 분석 결과를 도출한다. 6장에서 결론 및 향후 연구를 다룬다.

2. 관련연구

최근 몇 년 동안 많은 침입탐지 관련 연구들이 이상탐지에 초점을 맞추고 있다. 물론 새로운 공격을 탐지하는 효율성을 볼 때 이상탐지에 집중하는 것이 합리적이다. 이상탐지를 높이기 위해 네트워크 침입탐지 개선방법으로 기계 학습 및 데이터마이닝 기술이 널리 사용되고 있다. 이러한 기술을 통해 네트워크 트래픽 이상 탐지를 자동화할 수 있지만 연구목적으로 사용할 수 있는 게시된 데이터가 부족하다.

T. Janarthanan과 S. Zargari[1]은 비정상탐지에서 연구를 수행할 수 있는 새로운 데이터셋(NSL-KDD)를 제안하였고 고차원 특징분석이 가능한 향상된 UNSW-NB15 데이터셋을 제안하였다. C. Khammassi와 S. Krichen[2]은 Wrapper 방식을 채택하여 특징선택을 하였고 LR을 분류자로 사용하고 기능 검색 전략으로 GA 알고리즘을 사용하였다. N. Moustafa 와 J. Slay[3]는 Association Rule Mining 알고리즘을 이용하여 특징값의 중심점과 FAR을 줄일 수 있는 모델을 제안하였다. M. Kamarudin 등[4]은 Logitboost-based, LR 알고리즘을 이용하여 특징선택 절차에서 관련성이 떨어지거나 중복되는 속성들을 제거하여 앙상블 분류방법을 사용하는 비정상 기반 침입탐지시스템을 제안하였다. K. Mwitondi와 S. Zargari[5]는 Clustering 및 Cross-Validation을 이용하여 검색기반 침입탐지를 위한 데이터 흐름 적용방법을 제안하였다. M. Belouch 등[6]은 RepTree 알고리즘과 프로토콜 서브 세트 기반의 2단계 분류기를 제시하였다. S. Guha[7]는 MANN (multimodal artiical neural network) 알고리즘을 기반으로 사이버 시스템의 네트워크 데이터를 수집하고 CNN, elbow method와 *k*-means clustering algorithm에 의한 비감독 학습모델을 제안하였다. N. Moustafa 등[8]은 Dirichlet 혼합모델을 기반으로 네트워크 데이터를 스니핑하고 수집하며 데이터를 분석하고 필터링하여 의사결정 엔진의 성능을 향상시켰다. M. Idhammad 등[9]은 ANN에 기반한 DoS 탐지방법을 제안하였고 제안된 방법에서 FNN (Feed-Forward Neural Network)은 최소 자원사용으로 DoS 공격을 정확하게 검출하도록 최적화 하였다. Table 1에서 침입탐지에 관련된 연구들을 연도, 데이터셋 및 알고리즘 별로 구분하여 비교한다.

Table 1. Comparisons of related works

Author	Year	Dataset	Algorithms
T. Janarthanan and S. Zargari	2017	UNSW-NB15	DT, SVM, KNN
C. Khammassi and S. Krichen	2017	UNSW-NB15	GA, LR, DT C4.5, RF, NBTree
N. Moustafa and J. Slay	2015	UNSW-NB15	Association Rule Mining

M. Kamarudin et al.	2017	UNSW-NB15	Logitboosst-based, LR
K. Mwitondi and S. Zargari	2017	UNSW-NB15	Clustering, Cross-Validation
M. Belouch et al.	2017	UNSW-NB15	RepTree
S. Guha	2016	UNSW-NB15	multimodal artiical neural network
N. Moustafa et al.	2017	UNSW-NB15	Dirichlet Mixed Model
M. Idhammad et al.	2017	UNSW-NB15	FNN, CFS

3. 데이터셋

UNSW-NB15 침입탐지 데이터셋[10]은 Fig. 1의 테스트베드를 구성하여 생성된다. IXIA 트래픽 생성기는 세 개의 가상 서버로 구성된다. 서버1과 서버3은 트래픽의 정상적인 확산을 위해 구성되고 서버2는 네트워크 트래픽에서 비정상적인 또는 악의적인 활동을 만들어 낸다.

서버들 간의 내부 통신을 설정하고 공용 및 사설 네트워크 트래픽을 수집하면 IP 주소가 10.40.85.30 및 10.40.184.30 인 두 개의 가상 인터페이스가 있게 된다. 서버는 두 개의 라우터를 통해 호스트에 연결된다. 라우터1은 10.40.85.1 및 10.40.182.1 IP 주소로 구성되고, 라우터2는 10.40.184.1 및 10.40.183.1 IP 주소로 구성된다. 이 라우터들은 정상 및 비정상 트래픽이 모두 통과되도록 구성된 방화벽 장비에 연결된다. Tcpdump 도구는 시뮬레이션 가동 시간 동안 Pcap 파일을 캡처하기 위해 라우터1에 설치한다. 이 테스트베드의 목표는 IXIA 툴에서 시작되어 네트워크 노드(예: 서버 및 클라이언트)들로 퍼져나가는 정상 또는 비정상 트래픽을 수집하는 것이다.

IXIA 도구는 정상 트래픽에 덧붙여 공격 트래픽을 생성한다. 실제 공격 환경과 유사한 공격 트래픽 생성을 위해 공격 행위는 CVE (Common Vulnerabilities and Exposures) 사이트[11]로부터 생성된다.

IXIA 도구를 사용하여 첫번째 시뮬레이션에서 초당 1번의 공격이 포함되도록 구성하였고, 두 번째 시뮬레이션에서는 초당 10번의 공격이 포함되도록 구성하였다. 시뮬레이션 과정에서 캡처한 데이터는 각각 50GB이다.

4. 데이터 전처리 및 제안방법

UNSW_NB15 데이터셋 전처리에 관계형 데이터베이스 도구인 MYSQL을 사용하고, 특징선택을 위해 데이터 마이닝 도구인 WEKA (Waikato Environment for Knowledge Analysis)[12]를 사용한다.

본 연구에서는 희소 클래스의 분류 성능을 개선하는 특징선택을 하고자 한다. 제안 방법에서 희소클래스는 다른 클래스들에 비해 상대적으로 적은 인스턴스 수를 갖거나 낮은 분류 성능을 보이는 클래스를 기준으로 한다. 희소 클래스는 Reconnaissance(3), DoS(4), Worms(8), Backdoor(9) 및 Analysis(10)의 다섯 클래스로 정한다.

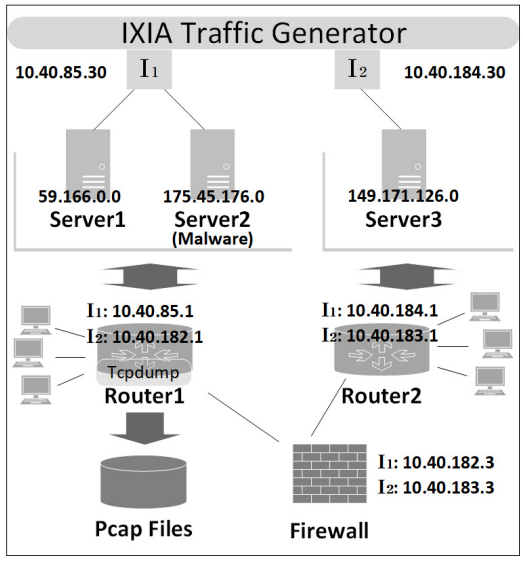

Fig. 1. UNSW-NB15 Testbed

Label cardinality of D is the average number of labels of the examples in D:

$$LC(D) = \frac{1}{|D|}\sum_{i=1}^{|D|}|Y_i|$$

$$imbalance\ ratio_i = \frac{|Y_i|}{LC(D) - |Y_i|}$$

Fig. 2. Formula of class imbalance ratio

Table 2와 Table 3은 Fig. 2의 수식 1에 의해 계산된 클래스 불균형 조절 전·후의 비율 및 인스턴스 수를 나타낸다. 상대적으로 클래스 불균형 비율이 높은 Normal과 Generic 클래스의 불균형 비율을 조절하여 특징선택을 하는데 소요되는 연산시간을 줄이고자 한다. 모든 클래스 중에서 Normal 및 Generic 클래스가 차지하는 비율의 순위를 바꾸지 않는 범위 내에서 언더샘플링(undersampling)[13]을 시도하였다.

Table 2. Class imbalance ratio for each class

Classes		Original imbalance ratio	Adjusted imbalance ratio
1	Normal	690.59%	2.81%
2	Exploits	1.78%	1.78%
3	Reconnaissance	0.55%	0.55%
4	DoS	0.65%	0.65%
5	Generic	9.27%	2.17%
6	Shellcode	0.06%	0.06%
7	Fuzzers	0.96%	0.96%
8	Worms	0.01%	0.01%
9	Backdoor	0.09%	0.09%
10	Analysis	0.11%	0.11%
Standard deviation		217.93%	1.00%

Table 3. The number of instances according to the class imbalance ratio

Classes		Instances	# of instances after undersampling
1	Normal	2,218,764	69,337
2	Exploits	44,525	44,525
3	Reconnaissance	13,987	13,987
4	DoS	16,353	16,353
5	Generic	215,481	53.871
6	Shellcode	1,511	1,511
7	Fuzzers	24,246	24,246
8	Worms	174	174
9	Backdoor	2,329	2,329
10	Analysis	2,677	2,677
Total		2,540,047	229,010

Table 4는 Normal 클래스의 인스턴스 수를 2^n분의 1로 줄여나가면서 각 클래스의 재현율(recall) 수치 변화를 보인다. Normal 클래스의 인스턴스 수가 줄어듦에 따라 가중 평균(weighted average)은 감소하지만 각 클래스에 대한 분류 성능의 변화는 적은 편이다. n이 5일 때 Normal 클래스의 불균형 비율은 2.81%로 낮아진다. Normal 클래스의 원래 인스턴스 수를 고려할 때 분류성능에 큰 영향을 주지 않으면서도 클래스 불균형은 상당히 완화된다.

Table 5는 Normal 클래스의 클래스 불균형을 조절한 후, Generic 클래스의 인스턴스 수 변화에 따른 재현율을 보인다. Generic 클래스의 경우에도 n이 2일 때 적절한 수준으로 클래스 불균형이 완화된다.

Table 4. Recalls according to the number of instances of Normal class

# / class	1/1	1/2	1/4	1/8	1/16	1/32
1	0.999	0.997	0.995	0.993	0.989	0.987
2	0.850	0.842	0.840	0.843	0.839	0.846
3	0.776	0.777	0.773	0.777	0.777	0.777
4	0.394	0.404	0.417	0.404	0.422	0.404
5	0.990	0.990	0.989	0.990	0.990	0.990
6	0.901	0.904	0.918	0.909	0.901	0.911
7	0.764	0.81	0.845	0.869	0.886	0.895
8	0.569	0.701	0.667	0.655	0.603	0.621
9	0.112	0.112	0.111	0.116	0.111	0.110
10	0.160	0.162	0.173	0.169	0.171	0.173
W. Avg.	0.986	0.976	0.962	0.947	0.933	0.923

Table 5. Recalls according to the number of instances of Generic class

# / class	1/1	1/2	1/4
1	0.999	0.997	0.995
2	0.850	0.842	0.84
3	0.776	0.777	0.773
4	0.394	0.404	0.417
5	0.990	0.990	0.989
6	0.901	0.904	0.918
7	0.764	0.810	0.845
8	0.569	0.701	0.667
9	0.112	0.112	0.111
10	0.160	0.162	0.173
W. Avg.	0.986	0.976	0.962

UNSW_NB15 데이터셋에서 제공하는 특징은 47개이다. 제안 방법에서는 srcIP와 dstIP 속성을 네트워크 클래스별로 분리하여 대상이 되는 특징을 Table 6에서와 같이 53개로 구성한다. Fig. 3은 실험에 사용하는 데이터 처리 및 실험 과정에 대한 흐름도이다. UNSW_NB15 데이터셋 중 Normal 클래스와 Generic 클래스에 대한 클래스 불균형 비율을 조절한 후, StratifiedRemoveFolds[12] 방법을 사용하여 훈련, 검증 및 테스트 데이터셋으로 나눈다. 훈련 및 검증 데이터와 래퍼 기반의 특징선택 알고리즘을 사용하여 모델 및 특징서브셋을 생성한 후, 모델에 테스트 데이터를 입력하여 결과를 도출한다. 이 중 가장 우수한 성능을 보이는 특징선택 기법을 제안한다.

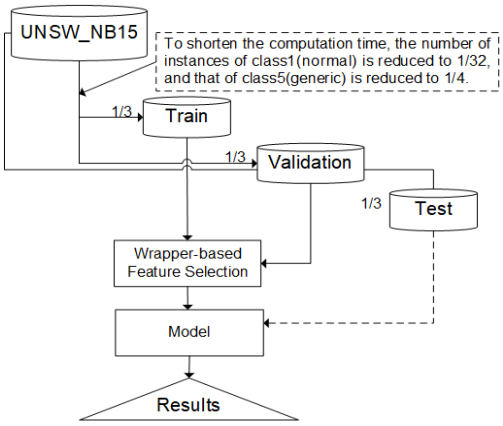

Fig. 3. Flowchart of data pre-processing and feature selection

Table 6. Features used in the proposed method

No.	Name	No.	Name
1	***srcip1***	28	dtcpb
2	***srcip2***	29	smeansz
3	***srcip3***	30	dmeansz
4	***srcip4***	31	trans_depth
5	sport	32	res_bdy_len
6	***dstip1***	33	Sjit
7	***dstip2***	34	Djit
8	***dstip3***	35	Stime
9	***dstip4***	36	Ltime
10	dsport	37	Sintpkt
11	proto	38	Dintpkt
12	state	39	tcprtt
13	dur	40	synack
14	sbytes	41	ackdat
15	dbytes	42	is_sm_ips_ports
16	sttl	43	ct_state_ttl
17	dttl	44	ct_flw_http_mthd
18	sloss	45	is_ftp_login
19	dloss	46	ct_ftp_cmd
20	service	47	ct_srv_src
21	Sload	48	ct_srv_dst
22	Dload	49	ct_dst_ltm
23	Spkts	50	ct_src_ ltm
24	Dpkts	51	ct_src_dport_ltm
25	swin	52	ct_dst_sport_ltm
26	dwin	53	ct_dst_src_ltm
27	stcpb	54	attack_cat (class)

특징선택은 크게 필터(filter) 기반의 방법과 래퍼(wrapper) 기반의 방법으로 구분할 수 있다. 제안 방법에서는 래퍼 기반의 방법을 사용하여 가장 우수한 성능을 보이는 특징 서브셋(subset)을 찾는다. WEKA의 특징선택 방법인 WrapperSubsetEval[12, 14]을 사용한다. 특징서브셋에 대한 평가 알고리즘은 SVM, k-NN, 의사결정트리(decision tree)를 사용하고, 탐색 알고리즘은 GA(genetic algorithm)[15], ANT(ant colony)[16], PSO(particle swarm optimization)[17]을 사용한다. 특징 서브셋에 대한 측정치(measure)는 정확도(accuracy)와 RMSE(root mean square error)를 사용한다. 특징선택에 사용한 데이터셋은 훈련(train)과 검증(validation) 데이터셋이다. 평가 알고리즘과 탐색 알고리즘의 다양한 조합을 통해 우수한 성능을 보이는 특징 서브셋들을 추출한다.

다양한 조합에 의해 추출된 특징 서브셋들은 분류 알고리즘을 사용하여 평가한 후, 희소 클래스의 분류에 우수한 성능을 보이는 특징추출기법을 제안한다.

5. 실험

UNSW_NB15 데이터셋의 특징 중 srcIP와 dstIP 속성을 네트워크 클래스 별로 구분하여 53개의 특징들로 구성한다. 54번 attack_cat 속성은 1-10의 값으로 클래스 구분에 사용한다.

No.	A	B	C	D	E	No.	A	B	C	D	E	No.	A	B	C	D	E
1						19						37					
2						20						38					
3						21						39					
4						22						40					
5						23						41					
6						24						42					
7						25						43					
8						26						44					
9						27						45					
10						28						46					
11						29						47					
12						30						48					
13						31						49					
14						32						50					
15						33						51					
16						34						52					
17						35						53					
18						36											

Fig. 4. Features selected by the wrapper-based experiments (A:SVM_GA, B:3NN_GA, C:DT_GA, D:DT_ANT, E:DT_PSO)

Fig. 4는 래퍼 기반의 특징선택 실험을 통해 다음 5가지 실험 조합으로부터 추출한 특징서브셋을 보인다. SVM_GA는 SVM 평가알고리즘과 GA 탐색을 사용한 실험이고, 3NN_GA는 k-NN 평가알고리즘과 GA 탐색을 사용한 실험이다. k는 3으로 설정한다. DT_GA는 의사결정트리 평가알고리즘과 GA 탐색을 사용한 실험이고, DT_ANT는 의사결정트리 평가알고리즘과 ANT 탐색 조합 실험이다. DT_PSO는 의사결정트리 평가알고리즘과 PSO 탐색 실험을 나타낸다.

Table 7. Comparison of classification performance of rare classes by feature subset (Recall)

Class	SVM_GA	3NN_GA	DT_GA	DT_ANT	DT_PSO
Normal	0.986	0.986	0.986	0.987	0.986
Exploits	0.869	0.953	0.910	0.853	0.809
Reconnaissance	0.783	0.773	0.780	0.780	0.787
DoS	0.258	0.098	0.208	0.335	0.459
Generic	0.981	0.982	0.982	0.983	0.982
Shellcode	0.911	0.907	0.893	0.885	0.927
Fuzzers	0.904	0.872	0.872	0.916	0.916
Worms	0.517	0.793	0.603	0.759	0.793
Backdoor	0.091	0.089	0.099	0.098	0.091
Analysis	0.139	0.135	0.158	0.186	0.164

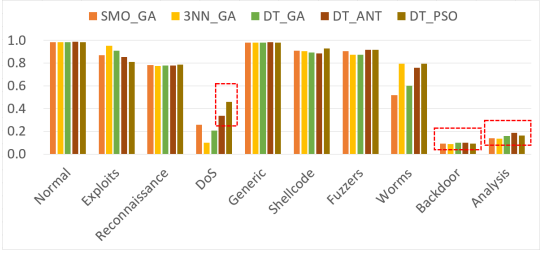

Fig. 5. Comparison of classification performance of rare classes by feature subset (Recall)

Table 8. Comparison of classification performance of rare classes by feature subset (ROC curve)

Class	SVM_GA	3NN_GA	DT_GA	DT_ANT	DT_PSO
Normal	0.999	0.999	0.999	0.999	0.999
Exploits	0.985	0.990	0.991	0.990	0.990
Reconnaissance	0.988	0.995	0.993	0.989	0.995
DoS	0.978	0.980	0.985	0.981	0.987
Generic	0.998	0.998	0.998	0.998	0.998
Shellcode	0.987	0.982	0.982	0.967	0.984
Fuzzers	0.991	0.993	0.993	0.993	0.993
Worms	0.896	0.948	0.896	0.940	0.957
Backdoor	0.974	0.979	0.984	0.979	0.977
Analysis	0.968	0.974	0.976	0.981	0.973

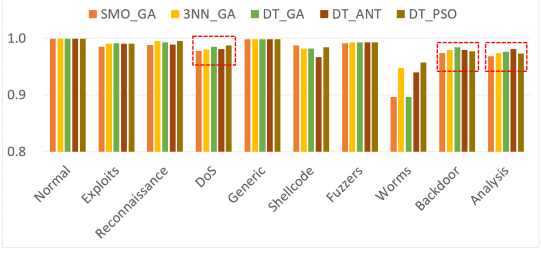

Fig. 6. Comparison of classification performance of rare classes by feature subset (ROC curve)

Table 9. Comparison of experimental results with other study

	SUBSET1		SUBSET2		
Class	TRAIN	TEST	TRAIN	TEST	DT_PSO
Normal	0.985	0.936	0.980	0.987	0.999
Exploits	0.943	0.873	0.952	0.950	0.990
Reconnaissance	0.929	0.814	0.978	0.977	0.995
DoS	0.909	0.854	0.917	0.911	0.987
Generic	0.990	0.990	0.995	0.996	0.998
Shellcode	0.777	0.731	0.937	0.996	0.984
Fuzzers	0.941	0.820	0.911	0.961	0.993
Worms	0.650	0.634	0.966	0.981	0.957
Backdoor	0.909	0.809	0.941	0.892	0.977
Analysis	0.937	0.830	0.955	0.913	0.973

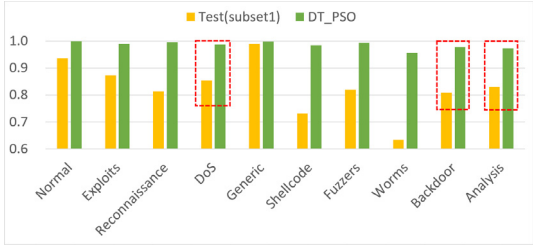

Fig. 7. Comparison of experimental results with other study (subset1, ROC curve)

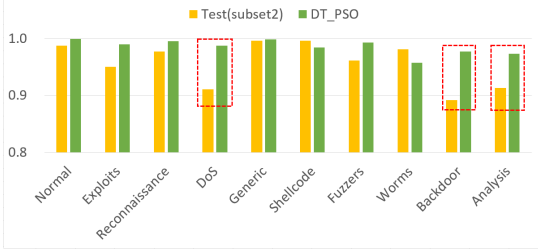

Fig. 8. Comparison of experimental results with other study (subset2, ROC curve)

Table 7과 Fig. 5는 다섯 가지 특징선택 실험 조합에 대한 비교 실험 결과를 재현율로 나타낸다. 이 중 DT_PSO 실험을 통해 추출한 특징서브셋이 전반적으로 우수한 성능을 보인다. DT_PSO 실험을 통해 선택된 특징 번호는 2, 3, 10, 14, 15, 16, 18, 20, 25, 29, 32, 35, 41, 44, 46이다. Table 8과 Fig. 6은 다섯 가지 특징선택 실험 조합에 대한 ROC 곡선 결과를 나타낸다.

Table 9와 Fig. 7-8은 기존의 다른 연구[1]와 제안 방법의 실험 결과를 비교한 ROC 곡선 결과이다. SUBSET1과 SUBSET2는 기존연구에서 제안한 특징선

택 실험 결과를 나타내고, DT_PSO는 본 연구에서 제안하는 방법의 실험 결과를 나타낸다. 실험결과는 제안 방법의 성능이 기존의 다른 연구에 비해 우수하거나 유사한 성능을 보인다.

6. 결론 및 향후연구

본 연구에서는 클래스 불균형의 정도가 심한 클래스에 대해 언더샘플링을 적용하여 연산시간을 단축시키고 래퍼 기반의 특징선택을 시도한다. 다양한 알고리즘을 사용하여 특징서브셋을 추출하고 가장 우수한 특징서브셋을 추출한 알고리즘을 제시하였다. 이 중 의사결정트리와 PSO 알고리즘을 사용한 특징서브셋이 가장 우수한 분류 성능을 보였다. 제안방법의 객관적 성능을 입증하기 위해 제안기법과 기존의 다른 연구[1]의 실험결과를 비교하였다.

향후, 제안방법에서 선택한 특징을 사용하여 희소클래스에 대한 성능을 높이는 연구를 하고자 한다. 희소클래스들에 적합한 데이터 전처리 알고리즘을 개발하고, 다양한 머신러닝 및 딥러닝 기법을 적용하여 희소클래스에 대한 분류 성능을 최대화한다.

REFERENCES

[1] T. Janarthanan & S. Zargari. (2017). Feature selection in UNSW-NB15 and KDDCUP'99 datasets. *In Industrial Electronics (ISIE), IEEE 26th International Symposium on*. (pp. 1881-1886). IEEE.

[2] C. Khammassi & S. Krichen. (2017). A GA-LR wrapper approach for feature selection in network intrusion detection. *computers & security, 70,* 255-277.

[3] N. Moustafa & J. Slay. (2015). A hybrid feature selection for network intrusion detection systems: Central points. *arXiv preprint arXiv:1707.05505*.

[4] M. Kamarudin, C. Maple, T. Watson, & N. Safa. (2017). A logitboost-based algorithm for detecting known and unknown web attacks. *IEEE Access, 5,* 26190-26200.

[5] K. Mwitondi & S. Zargari. (2017). A Repeated Sampling and Clustering Method for Intrusion Detection. *In International Conference in Data Mining (DMIN'17)*. (pp. 91-96). CSREA Press.

[6] M. Belouch, S. E. Hadai, & M. Idhammad. (2017). A two-stage classifier approach using reptree algorithm for network intrusion detection. *International Journal of Advanced Computer Science and Applications (ijacsa), 8(6)*, 389-394.

[7] S. Guha. (2016). *Attack detection for cyber systems and probabilistic state estimation in partially observable cyber environments*. Arizona State University.

[8] N. Moustafa, G. Creech & J. Slay. (2017). Novel geometric area analysis technique for anomaly detection using trapezoidal area estimation on large-scale networks. *IEEE Transactions on Big Data*.

[9] M. Idhammad, K. Afdel, & M. Belouch. (2017). Dos detection method based on artificial neural networks. *International Journal of Advanced Computer Science and Applications, 8(4)*, 465-471.

[10] *The UNSW-NB15 dataset.* (2018). www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets.

[11] *CVE (Common Vulnerabilities and Exposures).* (2018). cve.mitre.org.

[12] *WEKA.* (2018). www.cs.waikato.ac.nz/ml/weka.

[13] N. V. Chawla. (2009). Data mining for imbalanced datasets: An overview. *In Data mining and knowledge discovery handbook*. (pp. 875-886). Springer, Boston, MA.

[14] R. Kohavi & H. J. George. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97(1-2),* 273-324.

[15] J. rey Horn, N. Nafpliotis, & D. E. Goldberg. (1994). A niched Pareto genetic algorithm for multiobjective optimization. *In Proceedings of the first IEEE conference on evolutionary computation, IEEE world congress on computational intelligence,* (pp. 82-87).

[16] M. Dorigo, M. Birattari, C. Blum, M. Clerc, T. Stützle, & A. Winfield. (2008). Ant Colony Optimization and Swarm Intelligence. *The 6th International Conference, ANTS 2008,* Springer.

[17] Y. Shi. (2001). Particle swarm optimization: developments, applications and resources. *In evolutionary computation, 2001. Proceedings of the 2001 Congress on.* (pp. 81-86). IEEE.

이 대 범(Lee, Dae Bum) [정회원]

· 2011년 6월 : 국립 EARIST대학교 경영정보(경영정보학 석사)
· 2016년 2월 : 국립 EARIST대학교 경영정보(경영정보학 박사)
· 2016년 2월 ~ 현재 : 목원대학교 조교수
· 관심분야 : 정보보호, 인공지능
· E-Mail : dblee@mokwon.ac.kr

서 재 현(Seo, Jae Hyun) [정회원]

· 2008년 2월 : 광운대학교 컴퓨터과학과(공학석사)
· 2016년 2월 : 광운대학교 컴퓨터과학과(공학박사)
· 2017년 3월 ~ 현재 : 원광대학교 컴퓨터·소프트웨어공학과 교수
· 관심분야 : 최적화 알고리즘, 진화연산, 인공지능
· E-Mail : delphia7@wku.ac.kr