https://doi.org/10.33778/kcsa.2021.21.2.011

# Netflow를 활용한 대규모 서비스망 불법 접속 추적 모델 연구

이 택 현*, 박 원 형**, 국 광 호***


## 요 약

대다수의 기업은 유무형의 자산을 보호하기 위한 방안으로, IT서비스망에 다양한 보안 장비를 구축하여 정보보호 모니터링을 수행하고 있다. 그러나 서비스 망 고도화 및 확장 과정에서 보안 장비 투자와 보호해야 할 자산이 증가하면서 전체 서비스망에 대한 공격 노출 모니터링이 어려워지는 한계가 발생하고 있다. 이에 대응하기 위한 방안으로 외부자의 공격과 장비 불법 통신을 탐지할 수 있는 다양한 연구가 진행되었으나, 대규모 서비스망에 대한 효과적인 서비스 포트 오픈 감시 및 불법 통신 모니터링 체계 구축에 대한 연구는 미진한 편이다. 본 연구에서는 IT서비스망 전체 데이터 흐름의 관문이 되는 네트워크 백본 장비의 'Netflow 통계 정보'를 분석하여, 대규모 투자 없이 광범위한 서비스망의 정보 유출 및 불법 통신 시도를 감시할 수 있는 프레임워크를 제안한다. 주요 연구 성과로는 Netflow 데이터에서 운영 장비의 텔넷 서비스 오픈 여부를 6개의 ML 머신러닝 알고리즘으로 판별하여 분류 정확도 F1-Score 94%의 높은 성능을 검증하였으며, 피해 장비의 불법 통신 이력을 연관하여 추적할 수 있는 모형을 제안하였다.


# A Study on the Detection Model of Illegal Access to Large-scale Service Networks using Netflow

Taek-Hyun Lee*, WonHyung Park**, Kwang-Ho Kook***


## ABSTRACT

To protect tangible and intangible assets, most of the companies are conducting information protection monitoring by using various security equipment in the IT service network. As the security equipment that needs to be protected increases in the process of upgrading and expanding the service network, it is difficult to monitor the possible exposure to the attack for the entire service network. As a countermeasure to this, various studies have been conducted to detect external attacks and illegal communication of equipment, but studies on effective monitoring of the open service ports and construction of illegal communication monitoring system for large-scale service networks are insufficient. In this study, we propose a framework that can monitor information leakage and illegal communication attempts in a wide range of service networks without large-scale investment by analyzing 'Netflow statistical information' of backbone network equipment, which is the gateway to the entire data flow of the IT service network. By using machine learning algorithms to the Netfllow data, we could obtain the high classification accuracy of 94% in identifying whether the Telnet service port of operating equipment is open or not, and we could track the illegal communication of the damaged equipment by using the illegal communication history of the damaged equipment.


**Keywords : Netflow, Backdoor, Machine Learning, Fruad Detection, Intrusion Detection**

접수일(2021년 06월 25일), 게재확정일(2021년 6월 29일)

* 서울과학기술대학교 IT정책전문대학 산업정보시스템(주저자)
** 상명대학교 정보보안공학과(공동저자)
*** 서울과학기술대학교 기술경영융합대학 글로벌융합산업공학과(교신저자)

# 1. 서 론

최근 ICT 기술의 발전으로 인하여 다양한 사물들이 인터넷 서비스로 연결되는 만물인터넷(IoE, Internet of Everything)이 가속화되고 있다. 이러한 만물인터넷은 5G 서비스가 대중화되면서 더욱 광범위한 서비스 영역에 활용될 것으로 전망하고 있다[5]. 미국의 통신장비기업 CISCO에서는 인터넷에 연결되는 장치가 '18년 18억 개에서, '23년 140억 개로 약 7배 증가할 것으로 전망하고 있으며, 글로벌 리서치 전문기관인 가트너에 의하면 전 세계 사물인터넷(IoT, Internet of Things)의 개수가 '20년 240억 개로 예상한다.[11]. 이에 따라서 비즈니스 분야, 자동차분야, 가정, 헬스기기, 학습분야, 에너지 등 다양한 서비스가 인터넷에 연결될 것으로 전망하고 있다. 반면에 인터넷 기기에 대한 사이버 공격이 지속하여 증가하는 것으로 보고되고 있다. 미국의 보안 전문 기업 Sonicwall에 의하면, IoT 기기에 대한 공격 시도가 '19년 대비 '20년에 약 48% 증가하였으며, 보안이 취약한 IoT 기기 등을 통하여 기업의 주요 자산이 위험에 노출될 수 있음을 우려하였다[17]. 그 외에 장비 제조사의 백도어 문제와 같이 사업자가 통제하기 어려운 보안 문제들이 지속해서 발생하고 있다[7]. 이러한 사이버 위협에 대응하기 위해서 대다수의 기업은 IT서비스망에 다양한 보안 장비를 구축하여 유무형의 자산을 보호하기 위한 보안 모니터링을 수행하고 있다. 그러나 서비스망 고도화 및 확장 과정에 보안 장비 투자 및 보호해야 할 자산이 증가하면서 전체 서비스망에 대한 사이버 공격 및 정보 유출 모니터링이 어려운 한계가 발생하고 있다. 이러한 문제를 효과적으로 개선하기 위하여 다양한 연구가 진행되었으나, 대규모 서비스망에 대한 효과적인 외부 공격 노출 및 감시 체계 구축에 관한 연구는 미진한 편이다. 본 연구에서는 대규모 서비스망에서 수집하는 Netflow 통신 정보를 머신러닝 기법으로 분석하여 장비의 원격 접속 공격에 활용되는 Telnet 포트의 오픈 여부를 점검하고, 빅데이터 시각화 분석을 활용하여 장비의 불법 통신 이력을 추적할 수 있는 보안 모니터링 체계를 제안하였다. 논문의 구성은 다음과 같다. 2장은 관련 연구를 소개한다. 3장은 실험 분석, 4장은 실험 결과에 관해 설명한다. 5장은 결론을 요약하고 향후 연구 방향을 제시한다.

# 2. 관련 연구

## 2.1 사이버 보안 위협의 증가

정부 기관 혹은 제조사가 납품하는 장비에 특수한 목적으로 백도어를 설치하거나, 설치했다는 의혹이 제기되어 국가와 기업 간에 분쟁이 발생하고 있다. 또한, 장비의 자체적인 보안 취약점으로 인하여 지속적인 해킹 및 정보 유출 피해가 발생하고 있다. 대표적인 사례로 미국의 민간 기업인 Fortinet, CISCO, Juniper 등에서 장비 백도어가 확인되었으며[8][12], 중국의 경우에는 CCTV, 화웨이, ZTE 등에서도 장비 백도어 설치에 관한 문제가 발생하고 있다[7]. 관련하여, 미국에서는 '21 국방수권법(NDAA, National Defense Authorization Act)을 개정하여, 중국의 화웨이, ZTE 통신 장비를 직접적인 안보 위협 대상으로 지정하면서 국가 간 갈등이 고조되고 있다. 또한, IoT 기기에 대한 공격 시도가 '19년 1월~6월 1,347만 건에서 '20년 1월~6월 2,002만 건으로 약 48%가량이 증가하며, 전체 장비에 대한 사이버 위협이 증가하고 있다[17]. 이러한 사이버 보안 위협으로부터 기업을 보호하기 위해서 다양한 유무형의 투자를 진행하고 있다. 그러나 서비스망 구조가 복잡해지고, 소형화된 IoT 장비 사용이 증가하면서, 서비스 전반에 대한 비정상 행위를 감시하는데 한계가 발생하고 있다. 이에 따라서 대규모 서비스망에 대하여 효율적인 보안 감시 체계 구축 방안의 마련이 필요해지고 있다.

## 2.2 보안 이벤트 수집 기술 동향

대규모 서비스망에 대한 정보보호 상태를 감시하기 위해서 네트워크 트래픽 정보를 수집하여 분석에 활용할 수 있다. 트래픽 정보 수집 방법에는 패킷 정보 수집, 로그 정보수집, 플로우 정보 수집 방식이 있다. 패킷 정보 수집 방식은 통신 과정에

https://doi.org/10.33778/kcsa.2021.21.2.011

서 패킷 헤더와 페이로드 정보를 전수로 수집하여 상세한 통신 정보를 분석할 수 있지만, 수집 데이터가 방대할 경우에 분석 속도가 느린 단점이 있다. 또한 공격자가 악의적인 공격 행위를 은닉하기 위해서 SSH(Secure Sockets Layer), TLS(Transport Layer Security)와 같은 암호화 통신을 수행할 경우에 분석이 어려운 한계가 발생한다[17]. 이를 보완하기 위해서 암호화 통신을 복호화하여 분석할 수 있는 기술이 활용되고 있으나[6], 투자 비용 증가 등으로 인하여 대규모 서비스망과 소형 장비에 적용하기 어려운 문제가 발생하고 있다. 로그정보 수집 방식은 운영체제로그, 웹 로그, 보안 장비 로그 등 정형화된 로그를 수집하여 서비스의 보안 모니터링에 활용하는 방식이다. 단점으로 로그 포맷이 표준화되어 있지 않아서 별도의 데이터 가공 절차가 필요하며, 로그 정보 수집을 위해서 운영 장비에 별도의 연동 설정이 필요하다. 이 과정에서 운영 장비의 장애가 발생할 수 있으므로, Config 설정 과정에 주의해야 한다. 마지막으로 플로우 정보 수집 방식은 네트워크 통신 흐름을 관측하여 패킷의 통계 정보를 수집하는 방식으로, 대규모 서비스망을 모니터링하기 위하여 사용할 수 있다. 그러나 네트워크 통신에 대한 샘플링 통계 정보를 수집하여 활용하므로, 통신 데이터가 누락되는 문제가 발생할 수 있다. 또한, 전체 통신 과정에서 단편적인 요청과 응답의 통계 정보를 제공하므로, 의미가 적은 불필요한 데이터가 수집될 수 있는 문제가 존재한다.

### 2.3 Netflow 분석 기술 동향

대표적인 플로우 정보 수집 방법에는 CISCO사에서 개발한 Netflow 프로토콜 방식이 있다. Netflow는 네트워크 통신 흐름을 효과적으로 모니터링하기 위해서 개발된 네트워크 프로토콜로서, 주로 라우터와 스위치 장비를 통과하는 통신 흐름을 일정 기간 관측하여 패킷 통계 정보를 수집한다[4]. Netflow 프로토콜 버전은 V1~9까지 존재하며, IPv4를 지원하는 V5 버전을 가장 널리 사용하고 있다. 주요하게 수집하는 통신 정보에는 출발지 포트, 목적지 포트, 프로토콜 유형을 포함한 5-Tuple 정보가 있으며, 그 외 통신 사용량(Bytes Count), 패킷 개수(Packet Count), 통신 구간(Source AS Number) 등 다양한 정보를 제공한다[3]. 그 외에, 대규모 네트워크에서 생성되는 데이터를 수집하고 분석하기 위해서는, 효과적인 데이터 가공 및 저장 기법이 요구된다. 안혜선은 GPU 기반의 보안 이벤트 고속 필터링 기법을 제안하였으며[2], 강원철은 장기적인 Netflow 데이터 트래픽 분석을 위해서 Hadoop MapReduce 기반의 분석 방법을 제안하였다[1]. 본 연구에서는 대규모 서비스망에 대한 원격 접속 노출 및 비인가 통신에 대한 효과적인 보안 감시 체계를 구축하기 위해서 Netflow 통계 정보를 활용하였다. Netflow 정보는 네트워크 장비에서 손쉽게 추출할 수 있으며, 대규모 투자 없이 모니터링 범위를 확장할 수 있는 장점을 가진다. 또한, 대규모 데이터를 고속으로 가공하기 위해서 GPU DB를 활용하였으며, 장기간 데이터 보존 및 시각화 연동을 위해서 Hadoop 시스템을 활용하였다.

### 2.4 Netflow 정보 보안 활용 동향

대표적인 Netflow 데이터는 네트워크 통신 흐름에 대한 전반적인 분석을 수행할 수 있으므로, 다양한 분야에 활용되고 있다. 주요 활용 분야에는 네트워크 모니터링을 통하여 트래픽 볼륨 측정, 망 확장 계획 수립, DDoS공격 모니터링, 정보보호 감시 분야가 있다[10]. 정보보호 분야에서 비정상 접속을 탐지하기 위하여 활용된 연구들은 다음과 같다. Dias는 K-Means, DBSCAN, Agglomerative 3개의 알고리즘 결과를 합산하여 1개의 Score 값으로 생성하여, DDoS 공격 및 이상 접속을 탐지할 수 있는 알고리즘을 제안하였다. 또한, 입력 변수에서 Netflow 포트 번호를 System Port(0-1023), User Port(1024-49151), Dynamic Port(49152-65535)로 구분하여 연구를 수행하였다[14]. Najafabadi는 HDFS(Hadoop Distributed File System) 분석 환경을 구성하여, 운영 장비 원격 접속을 시도하는 SSH Bruteforce 공격을 Nearest Neighbor 알고리즘으로 탐지할 수 있는 기법을 제안하였다[15]. 또한, Proto는 대규모 네트워크에서 IQR(Interquartile Range) 통계 기법을 활용하여, Netflow

데이터 기반으로 이상 접속을 탐지할 수 있는 방법을 제안하였다[16]. 그 외에 Thapngam는 Pearson's의 상관계수분석기법으로[18], Hameed는 Counter-Based 알고리즘 분석 기법으로 DDoS 공격을 탐지하는 알고리즘을 제안하였다[13]. Bilge는 주요 변수 선택 및 머신러닝 지도학습 모델을 활용하여 Botnet 검출 방법을 제안하였으며[9], Zhang는 Netflow 정보를 시계열적으로 분석하여 이상 탐지 결과를 시각화하는 방법을 제안하였다[19]. 선행연구를 통하여 Netflow 정보가 다양한 정보보호 분야에 활용되는 것을 볼 수 있다. 즉, Netflow 데이터는 네트워크 장비에서 어렵지 않게 추출할 수 있으며, 통신 요약 정보를 제공하므로 다양한 정보보안 분야에서 활용하는 것을 볼 수 있다. 그러나 Netlfow 데이터를 머신러닝 기법으로 분석하여, 외부 공격의 기본 수단으로 활용되는 Telnet 서비스 포트의 오픈 여부를 식별하고, 비정상 통신을 감시할 수 있는 연구는 미진한 편이다. 따라서 본 연구에서는 Netflow 데이터를 머신러닝 기법으로 분석하여, Telnet 서비스 포트 오픈 여부를 감지할 수 있는 머신러닝 모델을 제안하였으며, SAS VA(Visual Analytics) 도구를 활용하여, 주요 공격자에 대하여 비인가 통신을 추적할 수 있는 시각화 방안을 제안하였다.

## 3. 제안하는 네트워크 데이터 실험 분석

### 3.1 제안하는 실험 환경 구성

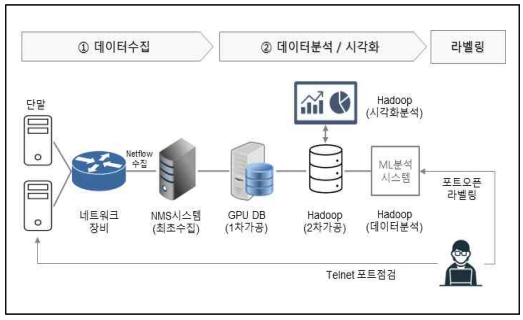
대규모 서비스망에서 수집한 Netflow 데이터를 분석하기 위하여 <표 1>과 같은 분석 환경을 구성하였다. 대용량의 Netflow 정보를 효과적으로 분석하기 위해서 GPU DB를 활용하여 고속으로 데이터 전처리를 수행하였으며, 데이터 보존 및 시각화 연동을 위해서 Hadoop 시스템에 저장하였다. 데이터 탐색 및 모델링 개발을 위해서 SAS EG/EM 솔루션과 Jupyter Notebook의 Python 프로그램을 활용하였으며, 최종적으로 SAS VA 솔루션으로 탐지 결과를 시각적으로 연계 분석할 수 있는 대시보드를 구성하였다.

<표 1> 분석환경

| 분류 | 소분류 | 내용 |
|---|---|---|
| 수집대상 | - | 네트워크 장비 Netflow 정보 |
| 데이터 가공 | 전처리 | GPU DB 기반 고속 전처리 |
| | 데이터보존 | Hadoop 기반 데이터 장기 저장 |
| 데이터 분석 | SAS | SAS EG, EM |
| | 분석도구 | Python 3.7 |
| 시각화 | SAS | SAS Visual Analytics |

### 3.2 제안하는 실험 과정

Netflow 데이터를 분석하기 위하여 데이터수집, 데이터분석/시각화, 라벨링 절차를 진행하였다. 전체적인 분석 절차는 (그림 1)과 같다.




(그림 1) 분석환경

첫 번째, 대규모 서비스망에 대하여 목적지 포트 23번인 Netflow 정보를 분석 대상으로 수집하였다. 총 159,965개의 Netflow 통신 이력에서, 1회 이상 통신 기록이 존재하는 고유한 장비는 33,747대로 분류되었다. 고려사항으로 Netflow 데이터는 목적지에 실제 장비가 존재하지 않더라도, 네트워크의 출발지 주소를 TCP Syn Flag, UDP 프로토콜에서 변조할 수 있으므로, 실제 통신하지 않은 장비가 검출될 수 있다. 이러한 한계점을 점검하고, 데이터의 현황 정보를 확인하기 위해서 장비별 1일간 수집한 패킷에서 소량으로 수집된 고유 장비에 대한 데이터 수집 빈도를 분석하였다. 장비별 1일간 수집된 데이터 빈도가 1개인 경우 61%, 2개인 경우 24%, 3개인 경우 8%, 4개 이상인 경우 7%로 확인되었으며, 세부 분석 현황은 <표 2>와 같다.

https://doi.org/10.33778/kcsa.2021.21.2.011

<표 2> Netflow 분석대상

| 목적지 포트 | 총 수집 데이터개수 | 장비 고유대수 | 장비별 데이터 빈도 | | | |
|---|---|---|---|---|---|---|
| | | | 1개 | 2개 | 3개 | 4개이상 |
| 23 | 159,965 | 33,747 | 61% | 24% | 8% | 7% |

두 번째로 분석 대상인 총 159,965개의 데이터에 대하여 속성 정보 73개 항목으로 분류하여 데이터 마트를 생성하였다. 데이터 속성 항목은 선행 연구와 Netflow 내부 데이터의 발생 빈도를 분석하여 정의하였다. 세부적으로, 목적지IP 주소는 Key 정보로써 장비별 통계 생성의 기준이 되며, 출발지 포트는 프로그램 수행 용도를 구분할 수 있도록 시스템포트, 사용자포트, 동적포트로 정의하였으며[14], 출발지IP의 목적지IP 접속 시도 개수를 정의하였다. 목적지 포트는 Telnet 포트 23번에 한정하였으며, 프로토콜은 TCP, UDP로 정의하였다. TCP Flag는 1번이라도 발생한 이력이 존재하는 항목을 반영하였으며, 출발지가 해외 접속인 경우도 특성 필드로 정의하였다. Netflow 특성 필드에 대한 세부 내용은 <표 3>과 같다.

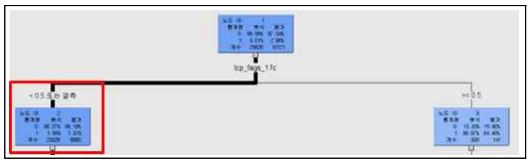
<표 3> Netflow 특성 필드(73개)

| 구분 | 설명 | 개수 |
|---|---|---|
| 수집시간 | 데이터를 수집한 시간(일별통계) | 1개 |
| 목적지IP | 목적지 IP주소 주소(Key값) | 1개 |
| 출발지IP | 목적지에 접속 시도하는 출발지 IP개수 | 1개 |
| 출발지포트 | 시스템포트 (0~1023) | 1개 |
| | 사용자포트 (1024-49151) | 1개 |
| | 동적포트(49152-65535) | 1개 |
| 목적지포트 | 원격접속 서비스(Telnet) 23번 포트 | 1개 |
| 프로토콜 | TCP(Proto 6), UDP(Proto 17) 프로토콜 | 2개 |
| TCP Flag | 0(NULL), 2(SYN), 4(RST), 16(ACK), 17(FIN-ACK), 18(SYN-ACK), 19(FIN-SYN-ACK), 20(RST-ACK), 24(PSH-ACK),25(FIN-PUSH-ACK), 26(SYN-PSH-ACK), 194(SYN-ECE-CWR) 등 총 59개 | 59개 |
| 해외유입 | 출발지 IP 주소가 해외 여부 점검(GeoIP활용) | 1개 |
| 탐지건수 | Netflow 정보 탐지 1일 건수 | 1개 |
| Packet합계 | 샘플링 패킷 개수 합계 | 1개 |
| Bytes합계 | 샘플링 데이터 총 Bytes 합계 | 1개 |
| 포트오픈 | NMAP 포트 점검 도구를 활용한 포트 오픈 점검 | 1개 |
| 총계 | | 73개 |

# 4. 제안하는 네트워크 데이터 실험 결과

## 4.1 머신러닝 수행 결과

장비에 대한 Telnet 서비스 포트 오픈 여부를 확인할 수 있는 머신러닝 모델을 개발하기 위해서 랜덤포레스트, 의사결정나무, 인공신경망, SVM, GLM, BN의 6개의 알고리즘을 사용하여 결과를 비교하였다. 학습 데이터와 검증 데이터를 각각 7대 3으로 분류하였으며, 모델에 대한 성능 측정은 Recall, Precision, Accuracy, F1-Score 지표를 활용하였다. 주요한 평가 지표로 장비의 실제 문제의 검출 정확도를 판별하는 Recall 지표와 모델이 탐지한 결과의 정확도를 판별하는 Precision 지표의 균형을 평가하는 F1-Score를 활용하였다. 머신러닝 분석은 Netflow로 생성한 데이터마트 전체를 사용하는 방법과 정확도를 높이기 위해서 Decision Tree 결과와 도메인 전문가의 판단 하에 모델에 사용하는 것이 불필요할 것으로 데이터를 제거하고, 모델을 수행하는 2가지 방법을 수행하였다. 첫 번째 방법으로 Netflow 데이터마트를 머신러닝 알고리즘으로 분석한 결과 가장 우수한 Decision Tree 알고리즘의 F1-Score가 70%로 정확도가 높지 않은 것으로 판단되었다. 이를 개선하기 위해서 Decision Tree에서 99.3%의 확률로 장비 포트가 오픈되지 않은 것으로 판단되는 Rule을 발굴하였으며, 이를 도메인 전문가와 검증하여 데이터 마트에서 제외하는 절차를 진행하였다. 적용된 필터링 Rule에 대하여 세부적으로 살펴보면 'TCP Flag 17(FIN-ACK) < 0.5 & TCP Flag 2(SYN) >= 0.5 & 해외유입 < 3.5'으로 정의할 수 있으며, 이는 TCP 프로토콜에서 연결을 종료하는 패킷이 0.5건 미만이고, 접속을 시도하는 SYN Flag가 0.5건 이상이며, 해외에서의 접근 시도횟수가 3.5건 미만인 경우는 불필요한 데이터로 판단하여 삭제하고 추가적인 분석을 진행하였다. SAS EM 도구에서 Decision Tree를 수행하여 분리 규칙을 생성하는 화면은 (그림 2)와 같다.



(그림 2) 의사결정나무 수행결과

두 번째 방법으로 데이터마트에서 학습에 활용하기 어려운 것으로 판단되는 데이터를 삭제하였으며, 이를 6개의 머신러닝 기법으로 분석한 결과 Random Forst와 Decision Tree 알고리즘의 F1-Score의 결과가 93%으로 높은 정확도를 제공하였다. 세부 실험 결과에 대한 내용은 <표 4>와 같다.

<표 4> 6개 모델 실험 결과

| 구분 (분석대상) | 알고리즘 | Precision | Recall | Accuracy | **F-Score** | **성능 순위** |
|---|---|---|---|---|---|---|
| 필터이전 | RF | 0.95 | 0.53 | 0.99 | 0.68 | 3 |
| | **DT** | **0.96** | **0.55** | **0.99** | **0.70** | **1** |
| | ANN | 0.85 | 0.58 | 0.98 | 0.69 | 2 |
| | SVM | 0.99 | 0.28 | 0.98 | 0.44 | 5 |
| | GLM | 0.86 | 0.40 | 0.98 | 0.54 | 4 |
| | BN | 0.81 | 0.41 | 0.98 | 0.54 | 4 |
| 필터이후 | **RF** | **0.94** | **0.93** | **0.93** | **0.93** | **1** |
| | **DT** | **0.94** | **0.93** | **0.93** | **0.93** | **1** |
| | ANN | 0.85 | 0.97 | 0.90 | 0.90 | 2 |
| | SVM | 0.94 | 0.54 | 0.75 | 0.69 | 5 |
| | GLM | 0.92 | 0.73 | 0.83 | 0.81 | 3 |
| | BN | 0.93 | 0.60 | 0.77 | 0.73 | 4 |

추가로 Random Forest 알고리즘의 분류 정확도를 더욱더 높이기 위한 실험을 진행하였다. 분석 데이터를 ① 원래데이터, ② 정규화, ③ 표준화하여 각각 실험 데이터를 생성하였다. 입력 값으로 활용되는 데이터를 정규화, 표준화하는 것은 머신러닝 모델 학습 과정에서 각 데이터가 유사한 영향도를 행사할 수 있도록 변환한 것이다. 또한, 총 73개의 변수 중 입력 값으로 사용되는 16개의 주요 변수를 선택하고, PCA 주성분 분석을 수행하여 입력 값의 편차를 줄이기 위한 기법을 적용하였다. 수행 결과로 학습 데이터를 정규화 하였을 경우에 F1-Score의 성능이 94%로 약 1%가량 향상되는 것을 확인할 수 있었다. <표 5>는 세부 실험 결과이다.

<표 5> Random Forest 실험결과

| 데이터유형 | Precision | Recall | Accuracy | **F-Score** | **성능순위** |
|---|---|---|---|---|---|
| ① 원래데이터 | 0.94 | 0.93 | 0.93 | 0.93 | 2 |
| **② 정규화** | **0.94** | **0.94** | **0.93** | **0.94** | **1** |
| ③ 표준화 | 0.93 | 0.93 | 0.93 | 0.93 | 2 |

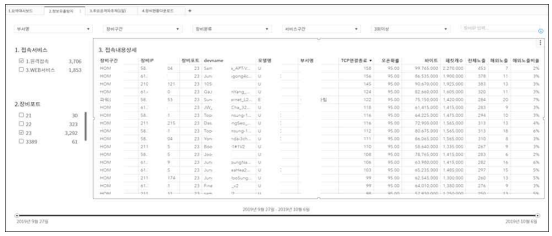
변수 선택 과정에서 주요 변수로 해외유입, TCP Flag 2(SYN), TCP Flag 17(FIN-ACK) 등 총 16개가 선택되었다. 이를 해석하면, 해외에서 접속이 발생하거나, TCP 프로토콜의 연결 시도와 종료 시도가 존재하는 경우에 운영 장비의 서비스 포트가 오픈되었을 확률이 높은 것으로 판단할 수 있다.

## 4.2 불법 통신 이력 추적

Netflow 통계 데이터의 장비 접속 이력과 원격 접속 포트 오픈 식별 머신러닝 모델을 활용하여, 외부 공격 노출 및 불법 접속 이력을 추적할 수 있는 감시 체계를 구축하였다. 본 연구에서는 SAS VA 도구를 활용하여, 시각화 및 연관분석 기능을 구현하였으며, 세부 분석 절차는 다음과 같다.

Step1) 서비스망에서 분석하고자 하는 대상을 선별하여 원격 접속이 시도된 장비를 정렬한다.
Step2) 머신러닝 분석 결과 Telnet 접속 포트가 오픈될 확률이 높은 장비를 선별한다.
Step3) Telnet 포트가 오픈되어 있는지 수동 점검한다.
Step4) 출발지 IP 주소의 Virustotal, Malwares.com, C-TAS에 유해성 여부를 점검한다.
Step5) 유해한 IP 접속 차단, 접속 소명을 진행한다.

(그림 3)은 불법 통신 이력 추적에 활용하는 대시보드 화면이다.



(그림 3) 점검 대시보드 화면

이를 통하여 Netflow 통신 데이터만을 활용하여 원격 접속 노출 추정 단말의 불법 통신 이력을 추적할 수 있는 활용 체계를 구축할 수 있었다. Net flow 데이터는 네트워크 장비에서 수집하므로, 공격자가 통신 이력을 회피하기 어려운 장점을 가지고 있다. 단점으로는 다양한 네트워크 데이터를 샘플링하여 수집하므로 부정확한 정보가 수집될 수 있으나, 기존의 수집 로그와 연계하여 정확도를 개선할 수 있을 것으로 판단된다.

## 5. 결론

본 연구에서는 Netflow 통계 정보를 활용하여, 원격 접속에 활용되는 Telnet 서비스 포트 오픈 여부 식별 및 불법 통신 이력을 시각적으로 추적할 수 있는 효과적인 방안을 제안하였다. 연구 결과에 대한 시사점은 다음과 같다.

첫 번째, Netflow 데이터에 대한 특징을 분류할 수 있는 73개의 항목을 제시하였다. 이를 응용하여, 추가적인 머신러닝 모델 개발 및 정보보안 활동에 활용할 수 있을 것으로 기대된다. 두 번째, Telnet 서비스 포트 오픈 점검을 별도의 보안 장비 투자 없이 안전하게 수행할 수 있는 머신러닝 모델을 제안하였다. 또한, 모델의 정확도 검증 결과 F1-Score 94%의 높은 성능을 검증하였다. 세 번째, 운영장비 포트 오픈을 판별할 수 있는 주요한 변수 16개를 발굴하였다. 분석 결과 Netflow 데이터에서 해외에서 접속을 시도하고, 접속 연결 및 종료한 패킷이 존재할 경우에 서비스 포트가 오픈되었을 가능성이 매우 높음을 확인하였다. 네 번째, 머신러닝 결과 장비 서비스 포트가 오픈된 것으로 추정되는 경우에, 추가로 주요 공격자를 추적하거나 정보 유출을 감시할 수 있는 프레임워크를 제안하였다. 이를 통하여 장비에 대한 외부 불법 접속 및 백도어에 의한 내부 정보 유출 시도를 추정할 수 있을 것으로 판단된다.

향후 추가 연구로는 Netflow는 네트워크의 통신 데이터를 Flag별 수집하므로, 실제 접속하지 않은 부정확한 정보가 저장될 수 있다. 이를 보완하기 위해서 장비 내부에 저장된 로그 정보와 연계 분석하여, 탐지 정확도를 높이는 연구가 필요하다. 또한, Telnet 서비스에 한정하여 포트 오픈 여부를 분석하였으나, 이를 확장하여 SSH, RDP등 다른 원격 프로토콜에 대한 확대 적용 검토가 필요하다.

## 참고문헌

[1] 강원철. "MapReduce 기반의 대용량 트래픽 분석 도구." 국내석사학위논문 忠南大學校 大學院, 2011.

[2] 안혜선, 박제원, 최재현, 이남용. "GPU기반의 보안 로그 이벤트 고속필터링기법에 대한 실증적 연구." 한국정보기술학회논문지 11.9 (2013): 133-141.

[3] 임익규, 안명수, 박성봉, 10기가급 패킷 캡쳐링에 의한 트래픽 분석 및 망 감시 시스템, KR101602189B1, 2015-04-28, 2016-03-11.

[4] 최상용, 천은영, 고대식. (2019). Suricata를 이용한 대용량 네트워크 트래픽 수집성능분석. 한국정보기술학회논문지, 17(8), 59-66.

[5] 탕천연. "디지털 시대의 포스터 디자인 문화와 발전 방향에 관한 연구." 국내박사학위논문 인천대학교 일반대학원, 2020.

[6] 한태현,이현명,조효재,조희승. "암호화 통신의 모니터링을 위한 SSLSPLIT 성능 분석." 정보과학회 컴퓨팅의 실제 논문지 25.10 (2019): 485-492.

[7] 화웨이/논란, 나무위키 홈페이지, 2021년03월08월 수정, 2021년03월08월 접속, https://namu.wiki/w/화웨이/논란.

[8] Backdoors Keep Appearing In Cisco's Routers, Tom's Hardware, 2018년 7월 19일 수정, 2021년 3월 8일 접속, https://www.tomshardware.com/news/cisco-back-door-hardcoded-accounts-software,37480.html.

[9] Bilge, Leyla & Balzarotti, Davide & Robertson, William & Kirda, Engin & Kruegel, Christopher. (2012). Disclosure: Detecting botnet command and control servers through large-scale NetFlow analysis. ACM International Conference Proceeding Series. 129-138.

[10] Cisco IOS NetFlow, CISCO 홈페이지, 2021년3월8월 접속, http://www.cisco.com/web/go/netflow.

[11] CISCO, Cisco Annual Internet Report (2018 - 2023) White Paper, March 2020.

[12] Fortinet Finds More SSH Backdoors, Bankinfo Security 홈페이지, 2016년 1월 25일 수정, 2021년 3월 8일 접속, https://www.bankinfosecurity.com/fortinet-finds-more-ssh-backdoors-a-8826.

[13] Hameed, S., Ali, U. HADEC: Hadoop-based live DDoS detection framework. EURASIP J. on Info. Security 2018, 11 (2018).

[14] L. Dias, S. Valente and M. Correia, "Go With the Flow: Clustering Dynamically-Defined NetFlow Features for Network Intrusion Detection with DynIDS," 2020 IEEE 19th International Symposium on Network Computing and Applications (NCA), Cambridge, MA, USA, 2020, pp. 1-10.

[15] M. M. Najafabadi, T. M. Khoshgoftaar, C. Calvert and C. Kemp, "Detection of SSH Brute Force Attacks Using Aggregated Netflow Data," 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA, 2015, pp. 283-288, doi: 10.1109/ICMLA.2015.20.

[16] Proto, André & Alexandre, Leandro & Batista, Maira & Oliveira, Isabela & Cansian, Adriano. (2010). Statistical Model Applied to NetFlow for Network Intrusion Detection. Transactions on Computational Science. 11. 179-191. 10.1007/978-3-642-17697-5_9.

[17] Sonicwall, 2020 Sonicwall Cyber Threat Report Mid-Year Update, July 2020.

[18] Thapngam T, Yu S, Zhou W, Makki S (2012) Distributed Denial of service (DDoS) detection by traffic pattern analysis. In: Peer-to-Peer networking and applications December 2014, Springer, Vol 7, Issue 4, pp 346 - 358

[19] Zhang, S., Shi, R. & Zhao, J. Seeflow: A Visualization System Using 2T Hybrid Graph for Characteristics Analysis of Abnormal Netflow. Wireless Pers Commun 101, 2127 - 2142

## 〔저자소개〕

이 택 현 (Taek-Hyun Lee)
2015년 서울과학기술대 산업정보시스템 석사
2021년 현재 서울과학기술대 산업정보시스템 박사과정
email : futp@naver.com

박 원 형 (Won-Hyung Park)
2002년 서울과학기술대 산업정보시스템 학사
2005년 서울과학기술대 정보산업공학과 석사
2009년 경기대학교 정보보호학 학사
2015년 성균관대학교 컴퓨터교육학 박사수료
2012년~2020년 극동대학교 사이버보안학과 부교수/학과장
현재 상명대학교 정보보안공학과 부교수
email : whpark@smu.ac.kr

국 광 호 (Kwang-Ho Kook)
1981년 서울대학교 산업공학과 석사
1989년 조지아공대 산업공학과 박사
2021년 현재 서울과학기술대학교 글로벌융합산업공학과 교수
email : khkook@seoultech.ac.kr