Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.32, NO.2, Apr. 2022 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2022.32.2.371

Gradient Boosting 기법을 활용한 다크넷 트래픽 탐지 및 분류

김 지 혜,[1†] 이 수 진[2‡]
[1]육군3사관학교 (강사), [2]국방대학교 (교수)

Darknet Traffic Detection and Classification Using Gradient Boosting Techniques

Jihye Kim,[1†] Soo Jin Lee[2‡]
[1]Korea Army Academy (Lecturer), [2]Korea National Defense University (Professor)

요 약

다크넷(Darknet)은 익명성과 보안을 바탕으로 하고 있어 각종 범죄 및 불법 활동에 지속적으로 악용되고 있으며, 이러한 오·남용을 막기 위해 다크넷 트래픽을 정확하게 탐지하고 분류하는 연구는 매우 중요하다. 본 논문에서는 그레디언트 부스팅 기법을 활용한 다크넷 트래픽 탐지 및 분류 기법을 제안하였다. CIC-Darknet2020 데이터셋에 XGBoost와 LightGBM 알고리즘을 적용한 결과, 99.99%의 탐지율과 99% 이상의 분류 성능을 나타내어 기존 연구에 비해 3% 이상 높은 탐지 성능과 13% 이상의 높은 분류 성능을 달성할 수 있었다. 특히, LightGBM 알고리즘의 경우, XGBoost보다 약 1.6배의 학습 시간과 10배의 하이퍼 파라미터 튜닝 실행시간을 단축하여 월등히 우수한 성능으로 다크넷 트래픽 탐지 및 분류를 수행하였다.

ABSTRACT

Darknet is based on the characteristics of anonymity and security, and this leads darknet to be continuously abused for various crimes and illegal activities. Therefore, it is very important to detect and classify darknet traffic to prevent the misuse and abuse of darknet. This work proposes a novel approach, which uses the Gradient Boosting techniques for darknet traffic detection and classification. XGBoost and LightGBM algorithm achieve detection accuracy of 99.99%, and classification accuracy of over 99%, which could get more than 3% higher detection accuracy and over 13% higher classification accuracy, compared to the previous research. In particular, LightGBM algorithm could detect and classify darknet traffic in a way that is superior to XGBoost by reducing the learning time by about 1.6 times and hyperparameter tuning time by more than 10 times.

Keywords: Darknet, Network Traffic Analysis, Gradient Boosting

I. 서 론

다크넷(Darknet)은 특정 소프트웨어, 설정, 권한 부여, 사용자 정의된 고유 프로토콜 등을 이용해야만 접속 가능한 형태의 오버레이 네트워크로서[1], 익명성과 보안을 바탕으로 하고 있어 악성코드 제작·배포, 정보 도청, 마약·무기 판매 등 다양한 범죄 및 불법 활동에 악용되고 있다. 2021년 5월 발생한 美 콜로니얼 파이프라인 랜섬웨어 감염 사태[2]와 프랑스 보험사 악사(AXA)에 대한 DDoS 공격[3]은 Tor 네트워크를 통해 발생하였고, 2022년 1월에는

Received(02. 03. 2022), Accepted(03. 10. 2022)
† 주저자, jseren@mnd.go.kr
‡ 교신저자, cyberkma@korea.kr (Corresponding author)

다크웹에서 유출된 7,971건의 카카오 계정을 악용한 대량의 크리덴셜 스터핑 공격이 발생하는 등[4] 다크넷을 활용한 범죄는 지속적으로 증가하고 있다.

다크넷에 접속하기 위해서는 주로 Tor 브라우저나 VPN을 이용한다. Tor는 다수의 중계서버(라우터)를 통해 랜덤화된 라우팅 경로를 이용하여 트래픽을 익명화하며, VPN은 사용자의 IP를 감추고 전송되는 데이터를 암호화하여 사용자의 개인 정보를 보호하는 기술이다. Tor 또는 VPN 트래픽 각각을 탐지하고 분석하는 연구는 활발하게 진행되어 왔으나[5-8], Tor와 VPN 트래픽을 동시에 다룬 연구는 상대적으로 많지 않다. 따라서, 다크넷의 오·남용 방지와 악성 활동의 근원지를 추적하기 위해 Tor와 VPN 전체를 다루는 다크넷 트래픽 탐지 및 분류에 대한 본 연구는 매우 중요하다고 할 수 있다.

그레디언트 부스팅은 여러 개의 약한 모델을 결합하고 분류 결과에 따라 가중치를 부여하여 강한 모델의 결과를 내는 앙상블 부스팅 기법에 속하며, 오류 그래디언트를 최소화하여 의사 결정 트리를 적합하게 만드는 경사 하강법(gradient descent)을 활용한다[9]. 본 논문에서는 그레디언트 부스팅 기반의 XGBoost와 LightGBM 알고리즘을 활용하여 다크넷 트래픽을 탐지 및 분류하였고, 두 알고리즘 모두에서 평균 99% 이상의 높은 정확도를 산출하였다.

본 논문의 나머지 구성은 다음과 같다. 2장에서는 Tor와 VPN 네트워크에 대한 개념 설명과 다크넷 트래픽과 관련된 기존 연구들을 소개하고, 3장에서는 본 연구에서 제안한 다크넷 트래픽 탐지 및 분류 방안에 대해 다룬다. 4장에서는 제안한 기법을 적용한 실험 결과와 평가를 제시하며, 마지막으로, 5장에서는 연구 결과를 요약하고 향후 연구를 제안하며 결론을 맺는다.

II. 배경지식 및 관련 연구

2.1 다크넷(Darknet)

인터넷은 Fig. 1.에 나타난 바와 같이 노출의 정도에 따라 크게 서피스 웹, 딥 웹, 다크넷(다크웹)으로 나눌 수 있다. 서피스 웹은 구글, 페이스북 등 일반적으로 접근할 수 있는 공개된 웹 서비스이다. 딥 웹은 일반 검색엔진에서 검색되지 않는 모든 웹 서비스를 말하며, 다크웹은 딥 웹의 범주에 속하지만 Tor 브라우저나 VPN과 같은 특수한 서비스를 이용해야

Fig. 1. Surface Web, Deep Web, Darknet[10]

만 접근이 가능한 웹 서비스를 말한다[10].

이러한 특성 때문에 다크넷은 흔히 각종 불법 온라인 활동의 진원지로 불리며, 다크넷을 통해 실행되는 응용 프로그램 또는 사용자 활동을 추적하기 위해서는 다크넷 트래픽 분석이 필수적이다[11].

2.1.1 Tor 네트워크

Tor(The Onion Router)는 TCP 기반의 익명 네트워크 웹 서비스로서, 물리 네트워크 위에 성립되는 가상의 네트워크인 오버레이 네트워크(overlay network)이다[12]. Tor는 중계 역할을 하는 약 6,000여개의 릴레이(프록시 서버)로 구성되어 있으며, 역할에 따라 크게 진입(entry guard relay) 노드, 중간 릴레이(middle relay) 노드, 최종 진출(exit relay) 노드로 나눌 수 있다[13].

Fig. 2. 와 같이 사용자는 Tor 브라우저를 이용하여 Tor 네트워크의 최초 진입 노드에 접속한 이후 중간 릴레이 노드를 통과하고, 마지막으로 최종 진출 노드를 거쳐 다크넷에 접속한다. 사용자의 트래픽은 각 노드의 공개키를 통해 암호화되며, 중간 및 진출

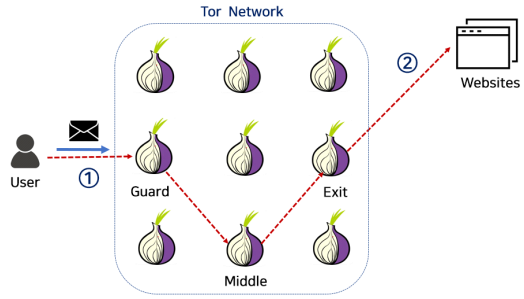

Fig. 2. The Structure of Tor Network

노드는 접속 시마다 임의로 변경된다. 따라서, 중간에서 트래픽을 가로챈다 해도 사용자의 트래픽이 어디에서 출발했는지, 최종 목적지가 어디인지 정확하게 알 수 없다.

또한, Tor를 통하는 모든 메시지는 “Cell”이라 불리는 동일한 길이(512Bytes)로 저장되기 때문에 다른 노드에서는 메시지의 내용이나 분량을 유추할 수 없으며, Tor는 위와 같은 방식으로 일반 웹 서비스보다 훨씬 높은 강도로 익명성과 프라이버시를 보장한다. 하지만 Fig. 2.의 사용자와 진입노드(Guard) 구간(①), 진출노드(Exit)와 웹사이트 구간(②)은 완전한 암호화가 되지 않아 스니핑 공격이 가능하다는 취약점이 존재한다. 본 논문에서 활용한 다크넷 데이터셋은 ① 구간에서 수집된 것이다.

2.1.2 VPN 네트워크

VPN(Virtual Private Network)은 글로벌 인터넷과 같은 공용 네트워크 인프라 내에 구성된 사설 네트워크로서[14], 두 엔드 포인트 간 완전한 암호화를 제공하여 사용자의 데이터가 외부에 노출되는 것을 방지하는 기술이다. VPN은 다크넷 접속뿐만 아니라 기업, 공공기관 내부망 접속 등 다양한 분야에서 활용되고 있다.

다크넷에서의 VPN은 주로 Tor 서비스가 금지된 국가에서 다크넷 접속을 해야 할 때나 Tor를 사용하고 있다는 사실이 탐지되는 것을 막아야 하는 경우에 사용되며, IP 주소를 숨기고 다른 국가의 서버를 통해 트래픽을 다른 경로로 우회하여 검열된 내용을 차단 해제하는 방법을 활용한다[15].

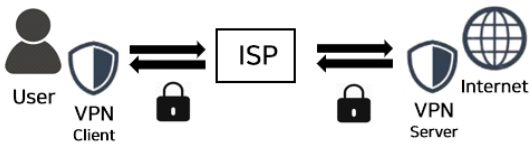

Fig. 3. The Structure of VPN

2.2 다크넷 트래픽 관련 연구

다크넷 트래픽 탐지 및 분류에 대한 연구는 주로 Tor 또는 VPN 트래픽 위주로 진행되어 왔다. A. H. Lashkari 등[5]은 수집된 Tor 트래픽의 시간 정보와 연관된 특성 23가지를 추출하여 KNN, Random Forest, C4.5 등의 머신러닝 기법을 적용하여 탐지 및 분류를 실시하였다. 그 결과, C4.5 알고리즘에서 평균 96.7%의 탐지 정확도와 Random Forest 알고리즘에서 평균 84%의 분류 정확도를 나타냈다. [6]에서는 동일한 방식으로 VPN 트래픽에 대한 탐지 및 분류를 시도하여, 평균 80% 이상의 정확도를 얻을 수 있었다.

본 연구에서 활용한 CIC-Darknet2020[16] 데이터셋은 [5]와 [6]에서 활용한 ISCXTor2016[17]과 ISCXVPN2016[18] 데이터셋을 결합한 형태로서, 총 8가지의 다양한 카테고리에서 Tor와 VPN 트래픽 모두를 다루고 있다. A. H. Lashkari 등 [19]은 CIC-Darknet2020 데이터셋을 활용하여 중요도가 높은 특성들을 추출한 뒤, 이미지 생성을 통해 2차원의 CNN 모델을 활용한 DeepImage 기법을 제안하였다. DeepImage 기법으로 다크넷 트래픽을 탐지 및 분류한 결과, 약 94%의 탐지 정확도와 약 86%의 분류 정확도를 달성할 수 있었다. 해당 연구는 Tor와 VPN 트래픽을 동시에 다루는 데이터셋을 최초로 생성하고, 딥러닝 기법을 활용하여 탐지 및 분류를 시행했다는 점에서 큰 의의가 있으나, 다중 분류에서의 분류 정확도가 카테고리 별로 약 48% 가량 차이나는 등 분류 성능이 일정하지 않다는 한계점을 가지고 있다.

M. B. Sarwar 등[11]은 CIC-Darknet2020 데이터셋을 활용하여 다크넷 트래픽을 탐지 및 분류하였다. 그들은 불균형한 데이터에서 데이터의 분포를 조정하는데 사용되는 오버샘플링 기법 중 하나인 SMOTE[20] 기법을 활용하여 데이터의 균형을 맞춘 이후, PCA, DT, XGB 등의 특성 추출 알고리즘과 다양한 머신러닝 및 딥러닝 모델을 적용하여 결과를 비교하였다. 그 결과, XGB 기법 특성 추출 및 CNN- LSTM 모델 적용 시 약 96.2%의 탐지 정확도 및 89%의 분류 정확도를 나타내어 기존의 연구[19]보다 향상된 성능을 도출할 수 있었다.

L. A. Iliadis 등[21]도 CIC-Darknet2020 데이터셋에 다양한 머신러닝 알고리즘을 적용하여 각각의 다크넷 탐지율을 비교하였다. 연구 결과, 다크넷 트래픽 탐지에서는 Decision Tree 알고리즘이 98.21%로 가장 높은 정확도를 보였고, Tor, Non-Tor, VPN, Non-VPN으로 나뉘어진 트래픽 분류에서는 Random Forest 알고리즘이 98.62%의 정확도로 가장 높은 성능을 나타냈다.

Neha Gupta 등[22]은 CIC-Darknet2020 데

이터셋에 XGBoost를 포함한 다양한 머신러닝 알고리즘을 적용하여 Normal, Tor, VPN 세 가지 형태로 트래픽을 분류하였고, 그 중에서 XGBoost 알고리즘이 약 98%로 분류 정확도가 가장 높았다. 해당 연구는 다크넷 트래픽을 Normal, Tor, VPN으로 나누어 높은 탐지율을 달성하기는 했으나, 검색, 채팅, 오디오, 비디오 등 다양한 카테고리에 대한 다중 분류를 실시하지는 않았다.

III. 다크넷 트래픽 탐지 및 분류 방안

3.1 제안 절차

본 논문에서 제안하는 절차는 아래 그림 Fig. 4.와 같다. CIC-Darknet2020 데이터셋을 활용하여 IP 위치 정보 추가 및 정규화 등 데이터 전처리를 실시한 이후 트래픽 탐지 및 분류를 위해 그레디언트 부스팅 기법인 XGBoost 및 LightGBM 알고리즘을 적용하였다. 이 때, GridSearchCV를 활용하여 하이퍼 파라미터를 최적화하였고, XGBoost와 Light GBM 알고리즘의 탐지, 분류 정확도 및 학습시간, 하이퍼 파라미터 튜닝 소요시간을 비교하였다.

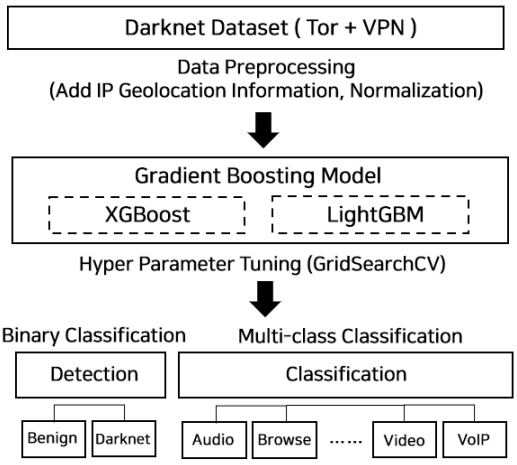

Fig. 4. A Proposed Methodology

3.2 데이터셋 특징

본 연구에서 활용한 CIC-Darknet2020 데이터셋은 총 8가지의 다양한 범주에서 Tor와 VPN 트래픽 모두를 다루고 있다. Table 1.에 나타난 바와 같이 총 158,659개의 데이터로 구성되어 있으며, 그 중 일반 트래픽은 134,348개(84.7%), 다크넷 트래픽은 24,311개(15.3%)이다. 다크넷 트래픽 중에서는 Audio-Stream이 13,284개(54.6%)로 가장 많고, P2P가 220개(1%)로 가장 적은 것을 확인할 수 있다.

Table 1. CIC-Darknet2020 Dataset

Category		# of data
Detection	Darknet	24,311
	Benign	134,348
Classification	Audio-Stream	13,284
	Browsing	263
	Chat	4,541
	File Transfer	2,610
	Mail	582
	P2P	220
	Video-Stream	1,346
	VoIP	1,465

3.3 데이터 전처리

CIC-Darknet2020 데이터셋은 네트워크 트래픽 흐름(Flow)을 생성하는 CICFlowMeter[23]을 활용하여 수집되었고, 각 데이터는 IP 주소, Flow의 지속시간, 포워드 패킷 개수 등을 포함한 총 86개의 통계적 정보들로 구성되어 있다. 본 연구에서 활용한 데이터 전처리 기법은 아래 Fig. 5.와 같이 크게 4단계로 나누어 설명할 수 있다.

먼저, IP 주소에 대한 세부 정보(지리적 위치, ASN/ISP 정보 등)를 얻을 수 있는 ipinfo 라이브러리[24]를 활용하였다. 데이터셋 각각의 출발지 및

① IP address Feature Extraction:
add 4 columns (src/dst IP geolocation, private src/dst IP (true/false))
* IP geolocation information (ipinfo library)

② Timestamp preprocessing:
Calendar Time (converts to) UNIX Time
ex. 2022-01-20 02:53:47 → 1642614827

③ Delete unnecessary information (3 columns) and missing value(NaN):
Flow ID, src/dst IP address

④ Data normalization (Sklearn StandardScaler)

Fig. 5. Data Preprocessing Procedure

목적지 IP 주소를 조회하여 해당 IP의 지역 정보와 사설 IP 여부를 확인하고, 이를 총 4개의 새로운 열에 저장하였다. 다음으로, 기존의 캘린더 형태로 저장되어 있던 시간 정보를 UNIX 시간 형태(1970년 1월 1일 00:00:00로부터 현재까지 누적된 초)로 변환하여 시간 정보에 대한 전처리를 실시하였다.

이후, 불필요 정보로 판단한 3개 열(Flow ID, 출발지 IP, 도착지 IP)을 삭제하고, 전체 데이터를 대상으로 결측값을 제거하였다. 마지막으로 ①-③번 절차를 통해 생성된 총 84개의 열을 대상으로 Sklearn StandardScaler 라이브러리를 활용하여 데이터 정규화를 실시하여 0~1사이의 값으로 변환하고, 학습(Train)과 테스트(Test) 데이터셋은 7:3의 비율(각 99036개, 42445개)로 구분하였다.

3.4 그레디언트 부스팅(Gradient Boosting)

그레디언트 부스팅은 여러 개의 약한 모델을 결합하여 더 강력한 모델을 만드는 앙상블 모델 중 하나로서 회귀, 분류, 순위 산출 등에 자주 활용된다[25]. 그레디언트 부스팅은 경사 하강법과 부스팅이 합쳐진 개념으로서, 여러 개의 모델이 순차적으로 학습을 진행하면서 이전 모델의 오차에 가중치를 부여하는 형태로 손실 함수를 최소화하면서 학습이 이루어진다.

본 논문에서 활용한 XGBoost 및 LightGBM 알고리즘은 부스팅 계열 알고리즘 중 초기 모델인 GBM 및 AdaBoost보다 발전된 형태로서 병렬 실행을 통해 학습 속도 및 과적합 제어 방식을 개선하였다[26]. 또한, 치우침이 있는 여러 모형의 평균을 통해 편차를 제거하고 분산을 감소시켜 일반적으로 더 뛰어난 예측 성능을 보이며, 표 형식의 정형화된 데이터(Tabular Data)에서는 딥러닝 모델(CNN, RNN)보다도 월등히 높은 성능을 산출하는 것으로 알려져 있다[27].

XGBoost는 새로운 트리가 기존 모델의 일부인 트리들의 에러를 수정하는 방식으로 학습을 수행하며[28], 성능 향상이 더 이상 나타나지 않을 때까지 새로운 트리를 계속해서 추가한다. LightGBM은 XGBoost보다도 더 발전한 형태의 모델로서 기존의 뛰어난 성능을 유지하면서도 학습에 걸리는 시간을 대폭 축소하였다. XGBoost와의 가장 큰 차이점은 균형트리 분할(Level Wise) 방식이 아닌 리프중심 트리 분할(Leaf Wise) 방식을 사용하여 트리가 수평적이 아닌 수직적으로 확장된다는 점이다. 최대 손실 값을 가지는 리프 노드를 지속적으로 분할하면서 트리의 깊이가 깊어지고 비대칭적인 트리를 생성하여 성능을 유지하면서도 균형 트리의 분할 방식보다 훨씬 더 빠르게 수렴한다는 장점을 가지고 있다[29].

XGBoost와 LightGBM의 하이퍼 파라미터 튜닝을 위해 Sklearn의 GridSearchCV 라이브러리를 활용하였으며, 과적합 제어를 위해 트리의 최대 깊이(max_depth) 값을 15로 설정하고, 트리 내 리프개수(num_leaves)는 2^{max_depth} 보다 작게 설정하였으며, 샘플링 비율(subsample) 값을 0.8로 설정하는 등 일부 파라미터들을 재조정하였다.

IV. 실험과 평가

4.1 실험환경

본 실험은 Ubuntu 20.04.3 LTS 운영체제, Intel(R) XEON Scalable Gold 6240 프로세서, 256GB RAM 사양의 RTX 3090 GPU 기반 워크스테이션에서 수행되었다. 또한, 학습시간의 비교를 위해 Google Colab Pro+의 Tesla P100-PCIE GPU 및 51GB RAM을 추가적으로 활용하였다.

4.2 실험결과 및 분석

4.2.1 다크넷 트래픽 탐지

테스트 데이터셋의 다크넷 트래픽 여부를 탐지한 결과, XGBoost와 LightGBM 모두 99.99%로 매우 높은 정확도를 나타냈다. Table 2.의 오차 행렬에 나타난 바와 같이 10개 이하의 소량의 오차 이외에는 모두 정확하게 탐지되었으며, 두 알고리즘의 성능은 큰 차이 없이 비슷하게 나타났다.

Table 2. Confusion Matrix (Detection)

True Label	Predicted Label			
	XGBoost		LightGBM	
	Benign	Darknet	Benign	Darknet
Benign	35128 (99.99%)	0	35128 (99.99%)	0
Darknet	10	7307 (99.99%)	6	7311 (99.99%)

Table 3. Comparison results on evaluation metrics of proposed approach and state-of-the-art study

Category		Detection (Accuracy)	Classification (F1 Score)							
			Audio	Browse	Chat	Email	File	P2P	Video	VoIP
Lashkari et al. [18]		0.94	0.92	0.51	0.88	0.67	0.75	0.93	0.85	0.59
M. B. Sarwar et al. [11]		0.96	0.82	0.85	0.83	0.84	0.85	0.94	0.86	0.87
Boosting Algorithm	XGBoost	**0.99**	**0.99**	**0.85**	**0.99**	**0.99**	**0.99**	**0.99**	**0.97**	**0.98**
	LightGBM	**0.99**	**0.99**	**0.94**	**0.99**	**0.99**	**0.99**	**0.99**	**0.98**	**0.98**

4.2.2 다크넷 트래픽 분류

테스트 데이터셋 트래픽을 대상으로 8가지 카테고리에 대해 분류를 실시한 결과는 아래 Table 3. 과 같다. 기존 연구 결과에 비해 정확도 및 F1-Score가 눈에 띄게 증가하였으며, 특히 LightGBM의 경우 데이터 수가 상대적으로 적었던 Browse와 P2P의 경우에도 높은 분류 성능을 유지하였다.

아래 Table 4.의 오차 행렬에서도 각 클래스에서 발생한 일부 오차 이외에는 대부분 정확하게 분류된 것을 볼 수 있다. 이를 통해 그레디언트 부스팅 기반 모델은 불균형 데이터셋에서 별도의 추가적인 샘플링 기법 없이도 일반적인 단일 모델보다 치우침 없이 뛰어난 예측 성능을 보이는 것을 알 수 있다.

Table 4. Confusion Matrix (Classification)
(XGBoost (Up) / LightGBM (Bottom))

True Label	Predicted Label							
	Audio	Browse	Chat	Mail	File Transfer	P2P	Video	VoIP
Audio	3949	2	23	0	0	0	1	0
Browse	6	64	2	0	0	0	5	3
Chat	2	0	1368	0	0	0	4	0
Mail	0	0	0	168	3	0	0	1
File Transfer	0	4	0	0	794	0	0	0
P2P	0	0	0	0	0	58	0	0
Video	0	0	0	0	0	0	426	4
VoIP	0	0	0	0	0	0	12	395

True Label	Predicted Label							
	Audio	Browse	Chat	Mail	File Transfer	P2P	Video	VoIP
Audio	3964	2	8	0	1	0	0	0
Browse	2	74	2	0	0	0	0	2
Chat	5	1	1363	0	2	0	0	3
Mail	0	0	0	169	3	0	0	0
File Transfer	1	0	0	0	794	0	2	1
P2P	0	0	0	0	0	58	0	0
Video	0	0	1	0	2	0	425	2
VoIP	3	0	0	0	0	0	9	395

4.2.3 소요시간 비교

Table 5. 에서는 XGBoost와 LightGBM 알고리즘에서의 학습 시간과 하이퍼 파라미터 튜닝 실행 시간을 비교하였다. Google Colab에서 두 알고리즘의 학습시간을 비교한 결과, LightGBM은 XGBoost보다 약 2.5배 빠른 속도를 나타냈다. 머신러닝 서버의 경우 뛰어난 성능 탓에 학습시간에서는 큰 차이를 볼 수 없었지만, 하이퍼 파라미터 튜닝 소요시간의 경우 약 11배의 차이를 나타냈다. 이를 통해 LightGBM 알고리즘이 높은 성능을 유지하면서도 XGBoost보다 훨씬 더 빠른 속도로 학습 및 하이퍼 파라미터 튜닝 등을 수행할 수 있다는 점을 직접 확인할 수 있었다.

Table 5. Comparison of Execution Time

	Category	XGB	LGBM
Google Colab	Learning Time	19.38s	7.76s
Server	Learning Time	2.45s	1.95s
	Hyper parameter Tuning Time	11833s	1057s

V. 결론 및 시사점

본 논문에서는 그레디언트 부스팅 기법 중 XGBoost, LightGBM 알고리즘을 활용하여 다크넷 트래픽을 기존 연구보다 더욱 빠르고 정확하게 탐지 및 분류하였다. IP 주소 및 시간정보 전처리, 결측값 제거, 데이터 정규화 등의 각종 전처리 기법과 XGBoost, LightGBM 알고리즘에 GridSearch CV를 활용한 하이퍼 파라미터 튜닝을 적용하여 다크넷 탐지 및 분류를 실시한 결과, 기존 연구에 비해 약 3% 이상의 높은 탐지 성능뿐만 아니라 데이터의 분포가 고르지 않은 불균형 데이터에서도 평균 99%

이상의 높은 분류 성능을 달성할 수 있었다. 특히, LightGBM의 경우 약 1.6배의 학습 시간과 약 10배의 하이퍼 파라미터 튜닝 실행 시간을 단축하여 탐지 및 분류의 효율성을 높일 수 있었다. 제안하는 부스팅 기법은 다크넷 범죄 수사 활동이나 조직에서의 보안 정책, 사이버 위협 관리 등에 보다 정확하고 적시적으로 활용할 수 있을 것으로 판단된다.

향후 연구에서는 Tor와 VPN 이외에도 다크넷 접속에 활용되는 보다 다양한 익명화 네트워크(I2P, Zeronet, Freenet 등)에서의 트래픽 패턴을 실제로 수집하여 비교해보려고 한다. 또한, CICFlow Meter라는 기존에 제작된 툴에 의해 추출된 패킷의 특성 이외에도 다양한 웹사이트 카테고리 별 또는 사용자 행위 별로 발생하는 다크넷 트래픽 특성들을 추출할 수 있는 자체적인 툴을 제작하는 것을 목표로 하고 있다.

References

[1] Jessica A. Wood, "The Darknet: A Digital Copyright Revolution," 16 Rich. J.L. & Tech 14, 2010.

[2] TrendMicro, "What We Know About the Darkside Ransomware and the US Pipeline Attack," https://www.trendmicro.com/en_id/research/21/e/what-we-know-about-darkside-ransomware-and-the-us-pipeline-attac.html/, accessed Jan 20, 2022.

[3] Cyberint, "Avaddon Ransomware Attack Hits AXA Philippines, Malaysia, Thailand and Hong Kong," https://cyberint.com/blog/research/avaddon-ransomware-attack-hits-axa-philippines-malaysia-thailand-and-hong-kong/, accessed Jan 20, 2022.

[4] Boannews, "A large-scale of login attempts happened by abusing the leaked Kakaotalk account," https://www.boannews.com/media/view.asp?idx=103990&kind=1&search=title&find=%C4%AB%C4%AB%BF%C0%C5%E5/, accessed Jan 10, 2022.

[5] Lashkari, A. H., Draper-Gil, G., Mamun, M. S. I., and Ghorbani, A. A, "Characterization of tor traffic using time based features," ICISSP, vol. 2, pp. 253-262. Feb, 2017.

[6] Draper-Gil, G., Lashkari, A. H., Mamun, M. S. I., & Ghorbani, A. A. "Characterization of encrypted and vpn traffic using time-related," ICISSP, vol.2, pp. 407-414. Feb, 2016.

[7] Florian Platzer, Marcel Schäfer, and Martin Steinebach, "Critical traffic analysis on the tor network," In Proceedings of the 15th International Conference on Availability, Reliability and Security (ARES '20). ACM, Article 77, pp. 1-10, Aug, 2020.

[8] P. Gao, G. Li, Y. Shi and Y. Wang, "VPN Traffic Classification Based on Payload Length Sequence," International Conference on Networking and Network Applications (NaNA), IEEE Computer Society. pp. 241-247, Dec, 2020.

[9] Machine Learning Mastery, "Gradient Boosting with Scikit-Learn, XGBoost, LightGBM, and CatBoost," https://machinelearningmastery.com/gradient-boosting-with-scikit-learn-xgboost-lightgbm-and-catboost/, accessed Jan 20, 2022.

[10] iTechHacks, "How To Use Deep/Dark Web On Your Android (A-Z Guide On Deep Web)," https://itechhacks.com/use-deep-web-on-your-android/, accessed Jan 10, 2022.

[11] M. B. Sarwar, M. K. Hanif, R. Talib, M. Younas and M. U. Sarwar, "DarkDetect: Darknet Traffic Detection and Categorization Using Modified Convolution-Long Short-Term Memory," IEEE Access, vol. 9, pp. 113705-113713, Aug, 2021.

[12] Dingledine, Roger, Nick Mathewson, and Paul Syverson, "Tor: The

second-generation onion router," Naval Research Lab Washington DC, 2004.
[13] Jihye Kim and Youngho Cho, "A Study of Tor Network Website Fingerprinting using Various ML Methods with PCA," Journal of Defense and Security, 3(2), pp 7-44. Dec, 2021.
[14] Ferguson, Paul, and Geoff Huston. "What is a VPN?." pp 1-22. 1998.
[15] VPNPROCLUB, "How To Use VPN For Dark Web," https://www.vpnproclub.com/how-to-use-vpn-for-dark-web/, accessed, Jan 10, 2022.
[16] UNB(University of New Brunswick), "CIC-Darknet2020", https://www.unb.ca/cic/datasets/darknet2020.html, accessed Dec 29, 2021.
[17] UNB(University of New Brunswick), "Tor-nonTor dataset (ISCXTor2016)", https://www.unb.ca/cic/datasets/tor.html, accessed Dec 29, 2021.
[18] UNB(University of New Brunswick), "VPN-nonVPN dataset (ISCXVPN2016)", https://www.unb.ca/cic/datasets/vpn.html, accessed Dec 29, 2021.
[19] Arash Habibi Lashkari, Gurdip Kaur, and Abir Rahali, "DIDarknet: A Contemporary Approach to Detect and Characterize the Darknet Traffic using Deep Image Learning," ICCNS 2020, ACM, USA, pp 1 - 13. Nov, 2020.
[20] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," J. Artif. Int. Res. 16, 1, pp. 321 - 357. Jan 2002.
[21] Iliadis, Lazaros Alexios, and Theodoros Kaifas, "Darknet Traffic Classification using Machine Learning Techniques," 10th International Conference on Modern Circuits and Systems Technologies (MOCAST). IEEE, pp. 1-4. July, 2021.
[22] Neha Gupta, Vinita Jindal and Punam Bedi, "Encrypted Traffic Classi fication Using eXtreme Gradient Boosting Algorithm," In International Conference on Innovative Computi ng and Communications. Advances in Intelligent Systems and Computing, vol 1394. Springer, pp. 225-232. Aug, 2019.
[23] Github, "CICFlowMeter," https://github.com/ahlashkari/CICFlowMeter/blob/master/ReadMe.txt, accessed Dec 29, 2021.
[24] Github, "ipinfo/python," https://github.com/ipinfo/python, accessed Jan 18, 2022.
[25] Li, Cheng, "A gentle introduction to gradient boosting," http://www.ccs.neu.edu/home/vip/teach/MLcourse/4_boosting/slides/gradient_boosting.pdf, accessed Jan 20, 2022.
[26] Towards Data Science, "How to choose between different Boosting Algorithms," https://towardsdatascience.com/how-to-select-between-boosting-algorithm-e8d1b15924f7, accessed Jan 20, 2022.
[27] Shwartz-Ziv, Ravid, and Amitai Armon, "Tabular data: Deep learning is not all you need," Information Fusion 81, pp. 84-90, Jun, 2021.
[28] Tianqi Chen and Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System," In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Association for Computing Machinery, USA, pp. 785 - 794, Jun, 2016.
[29] LightGBM, "LightGBM Features," https://lightgbm.readthedocs.io/en/latest/Features.html, accessed Jan 20, 2022.

〈저 자 소 개〉

김 지 혜 (Jihye Kim) 정회원
2013년 3월: 육군사관학교 응용물리학과 학사
2020년 3월: 국방대학교 컴퓨터공학과 석사
2021년 3월: 미국 해군대학원 컴퓨터과학 석사
2021년~현재: 육군3사관학교 컴퓨터과학과 강사
〈관심분야〉 네트워크 보안, 머신러닝

이 수 진 (Soo Jin Lee) 종신회원
1992년 3월: 육군사관학교 전산학과 학사
1996년 2월: 연세대학교 컴퓨터과학과 석사
2006년 2월: 한국과학기술원 전산학과 박사
2006년~현재: 국방대학교 국방과학학과 교수
〈관심분야〉 사이버전자전, 사이버안보, 인공지능 기반 사이버보안, 하드웨어 기반 사이버보안