https://doi.org/10.7840/kics.2022.47.11.1768

VPN/NoN-VPN 분류를 위한 트래픽 통계 정보의 중요도 분석

이 민 성◆, 박 지 태*, 백 의 준*, 최 정 우*, 김 명 섭°

Importance Analysis of Traffic Statistics Information for VPN/NoN-VPN Classification

Min-Seong Lee◆, Jee-Tae Park*, Ui-Jun Baek*, Jung-Woo Choi*, Myung-Sup Kim°

요 약

네트워크 환경이 급성장하면서 인터넷의 사용량이 증가하고 있다. 이러한 상황에서 네트워크 관리자의 효율적인 네트워크 운용이 필요해지고 네트워크 사용자는 적합한 QoS를 제공 받기 위해 정확한 트래픽 분류를 하는 것이 중요하다. VPN은 네트워크 통신의 보안 방법의 하나로 채택되고 있다. 하지만, 암호화의 악의적인 사용이 증가하고 방화벽을 회피하기 위한 도구로 활용되고 있다. 이에 따라 네트워크의 효율적인 운용을 위해 VPN과 NoN-VPN을 분류하는 방법이 필수적이다. 본 논문에서는 트래픽의 플로우 정보와 통계 항목들을 조합하여 VPN과 NoN-VPN을 분류하는 방법을 제안한다. 통계 정보들을 사용하였을 때 머신 러닝만을 활용하여도 약 98%의 분류 정확도를 보여줄 수 있으며, 플로우 정보와 통계 항목들중 분류에 중요한 역할을 하는 통계 정보를 분석한다.

키워드 : 트래픽 분류, 머신 러닝, 통계 정보, VPN, NoN-VPN
Key Words : Traffic Classification, Machine Learning, Statistical Information, VPN, NoN-VPN

ABSTRACT

With the rapid growth of the network environment, the usage of the Internet is increasing. In such a situation, efficient network operation of a network manager is required, and it is important for network users to accurately classify traffic in order to receive appropriate QoS. VPN is being adopted as one of the security methods of network communication. However, the malicious use of encryption is increasing and it is being used as a tool to evade firewalls. Accordingly, it is essential to classify VPNs and NoN-VPNs for efficient network operation. In this paper, we propose a method to classify VPN and NoN-VPN by combining traffic flow information and statistical items. When statistical information is used, classification accuracy of about 98% can be shown even by using only machine learning, and statistical information that plays an important role in classification among flow information and statistical items is analyzed.

※ 이 논문은 2021년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과(2021RIS-004)이고, 2022년도 정부의 재원으로 한국전자통신연구원의 지원을 받아 수행된 위탁연구개발과제의 결과(No.22RH1210, 이종산업간 5G/B5G 단대단 서비스 혁신과 제공 편리화를 위한 협력 테스트베드 핵심 기술)이다.
◆ First Author : Department of Computer and Information Science, Korea University, min0764@korea.ac.kr, 학생회원
° Corresponding Author : Department of Computer and Information Science, Korea University, tmskim@korea.ac.kr, 종신회원
* Department of Computer and Information Science, Korea University, {pjj5846, pb1069, choigoya97}@korea.ac.kr, 학생회원
논문번호 : 202208-169-B-RN, Received August 4, 2022; Revised September 7, 2022; Accepted September 9, 2022

Ⅰ. 서 론

네트워크 환경이 급성장하고 인터넷의 보급과 하드웨어 기술의 발전으로 인하여 인터넷 사용량이 증가하고 있다. 사회적인 영향으로 인해 재택근무가 증가하고 외부로의 출입이 적어지고, 변화하는 환경에 맞춰 다양한 응용 서비스들이 개발되면서 발생하는 트래픽의 양도 증가하고 있다. 이러한 상황에서 네트워크의 효율적인 운용 방식이 중요해지면서 네트워크 관리자는 서비스를 제공하면서 안정성과 신뢰성을 위한 네트워크의 효율적인 관리가 필요해지고, 사용자는 고품질의 서비스를 받는 환경이 필요하다. 네트워크 관리자는 네트워크 사용자에게 적합한 QoS(Quality of Service)를 제공해야 하며[1-3], 이를 위해 정확한 트래픽 분류를 하는 것이 중요하다.

트래픽 분류는 서비스를 실행하였을 때 서비스를 기반으로 트래픽의 플로우를 분류하는 작업이다. 트래픽의 플로우는 Source 및 Destination IP와 포트가 동일한 패킷들의 집합을 의미한다. 응용 트래픽 분야에서 다양한 응용 트래픽 분류 기법들이 사용되고 있다. 응용 트래픽을 분류하기 위한 기본적인 분류 방법은 포트 기반 분류 방법[4], 페이로드 기반 분류 방법[5], 통계정보 기반의 분류 방법[6]이 있다. 포트 기반 분류나 페이로드 기반 분류 방법은 구현이 간단하고 대규모 네트워크에서 매우 효율적이지만 암호화된 트래픽의 발생으로 인하여 응용 트래픽을 분류하기는 어려워지고 있다. 암호화된 트래픽을 분류하기 위하여 머신 러닝 및 딥 러닝 방법을 사용하여 분류하는 연구가 진행되고 있고, 플로우를 다양한 방법으로 전처리하여 각 모델에 입력하여 높은 분류 정확도를 가진 모델을 학습하기 위한 연구들이 많이 진행되고 있다.

VPN(Virtual Private Networks)은 네트워크의 통신 보안 방법의 하나로 채택되고 있다. 개인 통신을 위해 공용 네트워크에서 터널 기술을 사용하여 사설 네트워크를 구축하는 방법이다. VPN은 대규모 암호화된 연결을 허용하고 전송 시간을 단축하는 순기능이 있지만, 암호화의 악의적인 사용이 증가하고 VPN을 방화벽을 회피하기 위한 도구로 활용되고 있다. 이러한 잠재적인 보안 취약점을 해결하기 위하여 VPN을 관리 할 수 있어야 하며, VPN 트래픽과 NoN-VPN 트래픽을 분류 가능한 방법이 필요하다. VPN 트래픽을 따로 분류해 낼 수 있다면, VPN을 사용한 공격자의 공격을 사전에 대처할 수 있다.

본 논문에서는 VPN과 NoN-VPN을 분류하기 위해 트래픽 플로우의 통계정보를 사용한다. 기존 논문들에서 두 가지 클래스를 분류하기 위하여 트래픽에 여러 가지 기법들을 사용하여 머신 러닝 모델을 학습하였으며, 높은 분류 정확도를 보이기 위한 딥 러닝 모델들을 사용하였다. 하지만 여러 가지 기법들을 사용한 머신 러닝 모델들은 높은 분류 정확도를 보이지 않았다. 딥 러닝 모델들에서는 높은 분류 정확도를 보여주고 있지만, 학습 시간이 오래 걸리는 단점이 있다. 본 논문에서는 트래픽에 어떠한 기법을 사용하지 않고 트래픽 플로우의 통계정보만을 사용하여 머신 러닝의 앙상블 모델에서 높은 분류 정확도를 도출하였다. 또한, 플로우의 통계정보를 도출하는 패킷의 개수와 통계정보의 수를 증가시키며 실험을 진행하였고, 실험 결과를 바탕으로 VPN과 NoN-VPN을 분류하기 위하여 중요한 역할을 하는 통계정보를 분석한다.

본 논문은 서론에 이어 2장에서는 암호화된 트래픽 분류와 관련된 연구에 대하여 설명한다. 3장에서는 실험에 사용한 데이터셋의 설명과 트래픽 플로우에서 사용된 플로우 정보와 통계항목에 관해서 설명하고, 사용한 앙상블 알고리즘과 분류 시스템 설계에 관한 내용을 설명한다. 4장에서는 실험과 실험 결과에 대하여 설명하고 기존 연구들과의 결과 비교를 설명한다. 마지막으로 5장에서는 결론으로 제안하는 방법과 분류에 중요한 역할을 하는 통계정보 대한 내용과 향후 연구에 대하여 언급한다.

Ⅱ. 관련 연구

서론에서 언급한 바와 같이, 일반적인 트래픽 분류에서 사용하는 방법은 포트 기반 분류 방법과 페이로드 기반의 분류 방법이 있다. 포트 기반 분류 방법은 다양한 프로토콜을 사용하는 응용이나 둘 이상 혹은 임의의 포트를 설정할 수 있는 기능을 제공하는 서비스 등 구조가 복잡한 서비스에 대해서는 분류의 신뢰성을 가지기 어렵다. 페이로드 기반의 분류 방법은 분석률이나 분류 정확도의 측면에서 분석 성능이 높지만, 수작업을 통해 시그니처를 추출해야 한다. 새로운 서비스가 발생하게 되면 서비스를 분류하기 위하여 새롭게 시그니처를 추출해야 하며, 기존에 분류가 가능한 시그니처도 서비스의 업데이트에 따라 변경되어 서비스의 변화에 신속하게 대처해야 할 필요가 있다. DPI(Deep Packet Inspection)는 패킷 내 헤더 및 페이로드 정보를 모두 확인하여 서비스를 분류하는 패턴이나 시그니처를 찾는 방법이다. 거론된 세 가지 방법 모두 현재 많이 사용되고 있는 암호화된 트래픽에서는 적용하기 어렵다는 단점이 있다.

암호화된 트래픽에서 VPN과 NoN-VPN을 분류하기 위하여 머신 러닝 및 딥 러닝 방법을 사용한 다양한 연구들이 있다. VPN 트래픽을 비정상 트래픽으로 생각하고 비정상 트래픽 탐지에서 데이터 불균형 문제를 목표로 클러스터링을 기반으로 한 언더샘플링 방법을 사용하여 불균형 데이터 세트를 처리한 연구가 있다. 클러스터링 센터에서 가장 가까운 이웃 샘플을 샘플 포인트로 정하여 언더샘플링을 통해 가능한 대부분의 클래스에서 핵심 정보를 유지한다. AdaBoost 알고리즘과 K-RUSboost 알고리즘을 사용하여 최대 73%의 분류 정확도를 나타내었다[7].

딥 러닝을 기반으로 실시간 VPN을 탐지하기 위한 연구로 CNN(Convolutional Neural Network) 및 MLP(Multiple-Layer Perceptron) 모델을 사용하여 분류하는 방법을 제안한 연구가 있다. VPN과 NoN-VPN을 분류하여 99.87%의 Precision을 달성하였으며, 이는 딥 러닝 모델을 사용하여 높은 분류 정확도를 보일 수 있다는 것을 증명한다[8].

암호화 기술을 채택한 트래픽의 분류 기술로 Pruning 기반의 CNN을 사용한 PCNN을 적용한 연구가 있다. 암호화 트래픽의 클래스별 분류를 위하여 Pruning 방법을 사용하여 효율적인 암호화 트래픽 분류 방법을 제안하였다. CNN을 통해 자동으로 특징을 추출하고 Pruning을 적용하여 모델에 중요한 매개변수를 유지함으로써 모델의 크기와 계산이 줄어들고 성능에 영향을 주지 않는 방법을 사용하였다. 연구의 결과로 PCNN에서 94%의 Precision을 보였다[9].

VPN에 대한 네트워크 모니터링 및 사용자 액세스 제한을 위한 트래픽 조사를 위하여 VPN을 분류하고 탐지하는 다양한 알고리즘을 사용한 연구 결과를 보였다. 총 8가지의 머신 러닝 및 앙상블 모델을 비교하였고 RF(Random Forest)에서 93.8%의 가장 높은 분류 정확도를 보였다[10].

언급한 연구들은 트래픽 분류를 위하여 트래픽에 여러 전처리 방법들을 사용하였고, 높은 분류 정확도를 위하여 딥 러닝 모델을 사용하였다. 하지만, 딥 러닝 모델을 사용하기 어려운 환경이나 신속한 대처를 위한 곳에서는 경량화된 모델을 사용하는 것이 좋다.

Ⅲ. 본 론

본론에서는 논문에서 사용한 데이터셋의 정보에 대하여 설명하고 트래픽의 통계정보에 대하여 설명한다. 학습 모델에 입력으로 넣기 위하여 트래픽에 다양한 기법을 사용하는 것이 아닌 트래픽의 플로우 단위에서 뽑아낼 수 있는 통계정보만 사용한다. 플로우 정보와 통계항목의 조합을 적용하여 트래픽 플로우 단위의 통계정보를 추출한다. 다음으로는 실험에 사용된 앙상블 알고리즘을 설명하고 전체 시스템 개요에 대하여 설명한다.

3.1 데이터셋

본 항에서는 논문에서 사용한 데이터셋에 대하여 설명한다. 사용한 데이터셋은 트래픽의 다양한 분류를 실험 할 수 있는 공공데이터셋인 ISCXVPN2016 데이터셋이다. ISCXVPN2016 데이터셋은 VPN/NoN-VPN의 이진 분류, 응용 서비스별 분류, 어플리케이션 별 분류 등에 사용되는 공공 데이터셋이다. 해당 데이터셋에서 단일 패킷으로 적용되는 UDP 패킷을 제외하고 통계정보를 사용할 수 있는 TCP 플로우만 사용하였다. 데이터셋에 대한 정보는 표 1과 같다. 총 27811개의 TCP 플로우 개수 중 VPN에 해당하는 플로우의 수는 2076개이고 VPN이 아닌 플로우의 개수는 25735개이다. 실험에 사용되는 VPN의 플로우 개수가 VPN이 아닌 플로우 개수에 비해 적지만, 실제 환경에서 발생하는 네트워크 상황을 가정하고 실험을 진행하였다. 학습에 사용된 데이터셋의 개수는 VPN과 NoN-VPN을 포함한 22248개의 플로우를 학습에 사용하였고 교차 검증을 위한 검증 데이터의 개수는 5562개를 사용하였다. 학습된 모델을 테스트하기 위한 테스트 데이터셋은 5563개의 테스트 데이터셋을 사용하였다.

표 1. ISCXVPN2016 정보
Table 1. ISCXVPN2016 Information

ISCXVPN2016 Information	
VPN	2076
NoN-VPN	25735
Total	27811

3.2 통계정보 추출

플로우 통계정보는 플로우의 정보와 13가지 통계항목을 적용하여 통계정보를 추출한다. 플로우 내 패킷 크기와 패킷 간 도착 시간을 기준으로 하고 통계정보를 추출할 플로우의 첫 패킷의 개수를 정의하였을 때, 최소 78개의 통계정보를 추출할 수 있고, 최대 568개의 통계정보를 추출할 수 있다. 추출한 통계정보를 입력으로 앙상블 모델에 학습시켜 VPN과 NoN-VPN 트래픽을 분류한다.

3.2.1 플로우 정보

트래픽 플로우의 통계정보를 추출하기 전 통계정보와 조합하기 위한 플로우 정보는 표 2와 같다. 플로우 내 패킷들의 크기를 고려해야 하며, 패킷 간 도착 시간을 고려해야 한다. 플로우 내에서의 패킷의 크기나 시간 정보는 플로우를 분류하기 위한 중요한 역할을 하며 여러 연구에서도 기본적으로 사용되고 있다. 다음으로는 플로우 내에서 패킷들의 방향성을 고려한다. 특정 서비스를 분류할 때 패킷들의 방향성은 분류를 위한 정보가 되기도 한다. 플로우 내 Forward 패킷, Backward 패킷, 그리고 모두를 고려한 전체 패킷을 통계정보에 사용한다. 다음으로는 플로우 내 패킷을 얼마나 사용하여 통계정보를 추출할 것인지 결정한다. 플로우 내 패킷의 개수를 5개부터 30개까지 보고, 추가로 전체 패킷을 고려하여 통계정보 조합을 만들어 낸다.

표 2. 플로우 정보
Table 2. Flow Information

Flow Information	
Size	· Packet Size
Time	· Packet Inter Arrival Time
Direction	· Forward Packet · Backward Packet · ALL(Forward, Backward)
Num of Packet	· 5, 10, 15, 20, 25, 30, ALL

3.2.2 통계항목

플로우 정보와 조합할 통계항목들은 총 13가지이다. 13가지의 통계항목은 표 3과 같다. 플로우 내 패킷들의 크기와 패킷 간 도착 시간과 패킷의 크기를 기준으로 합, 최댓값, 최솟값을 사용하고 산술 평균, 기하 평균을 사용한다. 사용된 패킷의 백분위 수를 고려하게 되는데 백분위 수는 패킷을 4가지 동등한 구간으로 나누게 되고, 1분위 수부터 3분위 수까지 사용한다. 추가로 3분위 수에서 1분위 수를 뺀 값(IQR)도 사용한다. 마지막으로 분산 값, 표준 편차, 표본 표준 편차, 비대칭도, 첨도를 사용한다. 13가지 통계항목들은 트래픽 분류 분야에서 통계정보를 추출하기 위해 사용되는 항목들이다.

3.3 앙상블 알고리즘

논문에서 사용한 앙상블 알고리즘은 RF(Random Forest)와 GB(Gradient Boosting) 알고리즘을 사용하였다. 머신 러닝 알고리즘 중에서 앙상블 모델을 사용하는 것이 트래픽 분류 분야에서 높은 분류 정확도의 결과를 나타내고 있고, 딥 러닝과 비교하여 데이터의 처리 시간도 길지 않은 장점이 있다.

표 3. 통계항목
Table 3. Statistical Items

Statistical Items	
Stats	· Sum, max, min · Arithmetic mean, Geometric mean · Packet Quantile(First, Second, Third) · Inter Quantile Range(Q3 - Q1) · Variance · Population Standard Deviation · Sample Standard Deviation · Skewness · Kurtosis

RF[11]는 어렵지 않은 하이퍼 파라미터 조정과 높은 예측 및 분류 성능으로 인해 많은 연구에서 사용되었다. RF는 많은 수의 독립적인 의사 결정 트리를 형성하고 각 트리가 데이터에 대해 분류한 결과에서 투표를 시행하여 가장 많이 득표한 결과를 최종 분류 결과로 선택한다. RF에서 생성된 일부 트리는 오버피팅이 될 수 있지만 많은 수의 트리를 생성함으로써 오버피팅이 분류에 큰 영향을 미치지 않도록 한다.

GB[12]는 회귀 분석 및 분류 분석을 수행할 수 있는 부스팅 방법을 사용하는 앙상블 알고리즘이다. 하나의 초기 추정값을 평균으로 정하고 이전에 학습한 트리의 오류를 다음 트리의 학습에 영향을 준다. 예측한 값과 실제 값의 차이를 반영하여 새로운 트리를 생성한다. 새로운 트리를 생성하기 때문에 적절한 수의 트리를 선택하는 것이 매우 중요하다. 트리의 수가 많으면 오버피팅이 발생할 수 있고, 트리의 수를 낮게 설정하면 언더피팅이 발생할 수 있다.

3.4 분류 시스템 설계

본 항에서는 시스템 설계에 대하여 설명한다. 전처리 과정에서 추출한 플로우 정보 및 통계항목을 조합

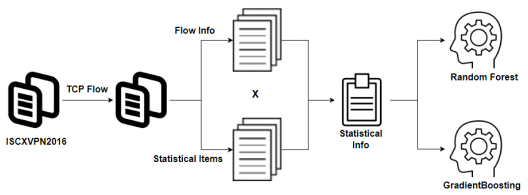


그림 1. 분류 시스템 개요
Fig. 1. Classification System Overview

하여 통계정보를 추출하였다. 실험은 크게 2가지로 진행하였다. 먼저, 플로우의 통계정보를 추출하는 패킷의 개수를 조절하여 최소 개수인 78개의 통계정보를 사용한 실험을 진행하였다. 두 번째 실험으로 패킷의 개수를 점점 늘려가면서 통계정보 78개부터 568개까지의 통계정보를 사용하였다. 그림 1은 시스템 설계에 대한 구성도이다.

Ⅳ. 실 험

본 장에서는 실험 결과에 대하여 설명한다. 통계정보 추출 시 플로우의 첫 n개의 패킷을 사용하여 실험 결과를 비교한 첫 번째 실험과 통계정보를 늘려가면서 분류 결과를 비교한 두 번째 실험을 진행하였다. 두 가지 실험을 진행하면서 트래픽에서 추출한 통계정보의 중요도를 비교하고 플로우 정보와 통계항목에서 VPN/NoN-VPN 트래픽 분류 시 중요한 역할을 하는 통계정보를 도출한다.

4.1 플로우의 첫 n개 패킷의 통계정보 실험

첫 번째 실험은 플로우 정보 중 통계항목을 적용할 패킷의 개수를 다르게 하여 실험을 진행하였다. 패킷의 개수는 5개부터 30개까지의 통계정보를 사용하는 것을 시작으로 전체 패킷의 통계정보를 고려하는 것으로 총 7번의 실험을 진행하였다. 표 4는 실험 결과를 보여준다. 앙상블 모델의 입력으로 78개의 통계정보가 입력으로 사용되었다. 패킷의 개수를 나누어 실험을 진행하였을 때 RF와 GB 모두 약 98%의 검증 정확도와 테스트 정확도로 수렴하였다. RF의 경우 모든 패킷을 사용한 통계정보를 적용하였을 때 가장 높은 분류 정확도를 보여주었고, GB도 역시 모든 패킷을 사용한 통계정보를 사용하였을 때 가장 높은 분류 정확도를 보였다. 플로우에서 가장 많은 패킷의 정보를 가지고 있을 때 VPN과 NoN-VPN을 가장 잘 분류할 수 있었다.

표 4. 플로우의 첫 n개 패킷의 통계 정보별 실험 결과
Table 4. Experimental results by statistical information of the first n packets of the flow

Num of Packet	Model	Val_Acc	Test_Acc
5	RF	98.22%	98.11%
	GB	98.29%	98.13%
10	RF	98.07%	98.14%
	GB	98.23%	98.45%
15	RF	98.11%	98.04%
	GB	98.35%	98.41%
20	RF	98.10%	98.11%
	GB	98.11%	98.41%
25	RF	98.41%	98.20%
	GB	98.20%	98.59%
30	RF	98.07%	98.20%
	GB	98.20%	98.54%
ALL	RF	98.30%	**98.41%**
	GB	98.61%	**98.63%**

4.2 통계정보 개수 별 실험

두 번째 실험은 통계정보의 개수를 다르게 하여 진행한 실험이다. 패킷의 개수를 기준으로 추출된 통계정보를 추가하여 실험을 진행하였다. 표 5는 두 번째 실험 결과를 보여준다. 실험 결과로 RF와 GB 모두 약 98%에 수렴하는 분류 정확도를 보여준다. 첫 번째 실험과 유사하게 통계정보를 가장 많이 사용하였을 때 높은 분류 정확도를 나타내었다.

4.3 통계정보의 중요도

두 번째 실험에서 도출된 학습 모델을 바탕으로 통계정보의 중요도를 측정하였다. 그림 2와 그림 3은 통계정보 546개를 사용하였을 때의 통계정보의 중요도의 상위 5개를 보여준 그림이다. 두 앙상블 모델에서 공통적으로 패킷 간 도착 시간에 대한 데이터가 높은 중요도를 보이며, 5개의 패킷에서 추출된 통계 데이터가 VPN과 NoN-VPN을 분류하는 중요한 역할을 한다는 것을 볼 수 있다. 마지막으로 통계항목 중에서는 백분위 수에 대한 값을 가장 많이 활용된다는 것을 알 수

표 5. 통계 정보 개수 별 실험 결과
Table 5. Results by number of statistical information

Num of Stats	Num of Packets	Model	Val_Acc	Test_Acc
78	ALL	RF	98.30%	98.41%
		GB	98.61%	98.63%
156	5, 10	RF	98.37%	98.23%
		GB	98.49%	98.47%
234	5, 10, 15	RF	98.44%	98.45%
		GB	98.61%	98.74%
312	5, 10, 15, 20	RF	98.44%	98.45%
		GB	98.63%	98.57%
390	5, 10, 15, 20, 25	RF	98.51%	98.50%
		GB	98.62%	98.57%
468	5, 10, 15, 20, 25, 30	RF	98.50%	98.50%
		GB	98.63%	98.66%
546	5, 10, 15, 20, 25, 30, ALL	RF	98.60%	**98.66%**
		GB	98.69%	**98.68%**

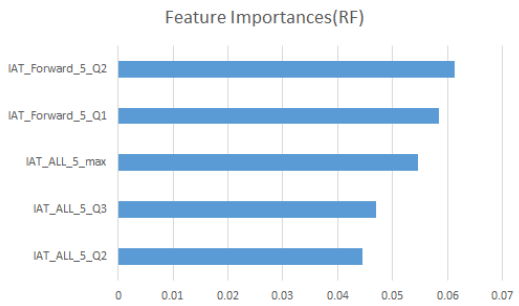


그림 2. Random Forest에서의 Feature 중요도
Fig. 2. Feature Importances in Random Forest

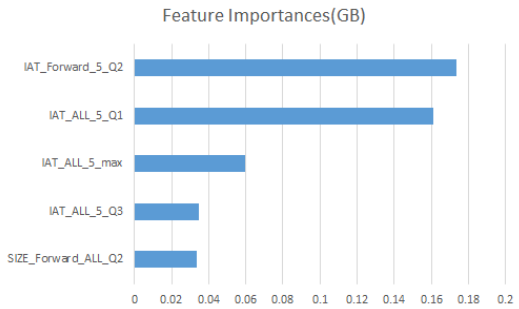


그림 3. GradientBoosting에서의 Feature 중요도
Fig. 3. Feature Importances in GradientBoosting

있다.

4.4 실험 결과

첫 번째 실험 결과 플로우의 첫 n개의 패킷을 사용하여 통계정보를 추출하였을 때 모든 패킷을 사용하여 통계정보를 추출하여 머신 러닝 알고리즘에 학습하였을 때 가장 높은 분류 정확도를 보였다. 모든 패킷을 고려하였을 때 VPN을 분류할 수 있는 정보의 양이 많기 때문이다. 하지만 전체 실험 결과에서 플로우의 첫 5개~30개의 패킷을 보는 것과 전체 패킷의 통계정보를 사용하였을 때 약 98%의 분류 정확도로 차이가 크지 않았다.

두 번째 실험 결과 통계정보의 수를 늘려가며 실험을 진행하여도 모든 실험에서 약 98%의 분류 정확도를 보이고 있으며, 가장 많은 546개의 통계정보를 사용하였을 때 높은 분류 정확도를 보였다.

마지막으로 통계정보의 중요도를 측정해 보았을 때, 플로우 정보에서 패킷 간 도착 시간의 정보와 플로우의 첫 5개의 패킷을 고려하였을 때 분류에 중요한 역할을 하고, 5개의 통계정보는 패킷의 백분위를 이용하는 것이 VPN/NoN-VPN 분류에 있어서 높은 영향을 준다는 것을 알 수 있다. 트래픽은 시간 정보를 가지고 있는 데이터이기 때문에 패킷 간 도착 시간의 정보가 분류에 중요한 역할을 하고 있다. 또한, 플로우나 패킷 단위에서 먼저 발생한 데이터가 트래픽을 분류하는 중요한 정보를 가지고 있다.

표 6. ISCXVPN2016 데이터셋의 기존 연구와의 비교
Table 6. Comparison with previous Papers of the ISCXVPN2016 Dataset

Model	Precision	Model	Precision
CART[6]	77.4%	GB[10]	80.1%
AdaBoost[6]	86.5%	RF[10]	91.8%
CNN[7]	99.87%	RF(Ours)	**98.66%**
PCNN[8]	94%	GB(Ours)	**98.68%**

ISCXVPN2016 데이터셋의 VPN/NoN-VPN 분류를 진행한 기존 연구와 비교해 보았을 때, 제안한 분류 시스템의 실험 결과가 좋다는 것을 알 수 있다. 기존 연구와의 분류 정확도를 비교한 것은 표6과 같다. 앙상블 모델 2가지를 사용하여 약 98%의 분류 정확도를 보여주고 있다. 머신 러닝 기반의 VPN/NoN-VPN을 분류하는 연구들에서는 RF에서 93.8%의 분류 정확도, GB에서 83.5%의 정확도를 보여주고 있지만, 통계정보를 활용하여 분류하였을 때 더 높은 분류 정확도를 도출하였다. CNN 및 MLP모델을 사용하여 99%의 분류 정확도를 보여주고 있지만, 머신 러닝과 딥 러닝 모델의 학습 시간을 고려하고, 분류 성능이 1% 정도의 차이가 있다. 이는 VPN을 분류하는 환경에 따라서 분류 시간과 학습 성능을 조율하여 적절한 모델을 사용할 수 있다.

V. 결 론

본 논문에서는 암호화된 트래픽에서 VPN과 NoN-VPN을 분류하기 위한 실험 방법과 사용된 통계정보에 대한 실험 결과를 통해 VPN/NoN-VPN 트래픽 분류에 중요한 플로우 정보와 통계항목에 대하여 결과를 도출하였다.

본 논문에서는 패킷의 개수 별, 통계정보 개수 별 분류 정확도를 비교하는 실험을 진행하였고, 모든 분류에서 98%에 해당하는 분류 정확도를 보여주었다. 이는 트래픽에서 다른 기법을 사용하지 않아도 통계정보만을 사용하여 VPN/NoN-VPN의 분류가 가능하다는 것을 보여준다. 사용된 패킷의 개수나 통계정보가 많을수록 차이는 적지만 더 높은 분류 정확도를 보여준다. VPN/NoN-VPN 트래픽을 분류할 때 많은 통계정보 중에서 플로우 정보 중 패킷 간 도착 시간과 플로우 내

5개의 패킷이 분류에 중요한 역할을 하고 있다. 통계항목에서는 패킷의 백분위 수를 고려하는 것이 분류에 큰 역할을 한다는 것을 보여주고 있다.

향후 연구로는 데이터셋에서의 서비스를 대상으로 분류를 진행할 예정이며, 각 서비스 내에서 VPN과 NoN-VPN을 분류하는 방법에 대하여 실험 및 연구를 진행할 예정이다.

References

[1] M.-S. Kim, Y. J. Won, and J. W.-K. Hong, "Application-level traffic monitoring and an analysis on IP networks," *ETRI J.*, vol. 27, pp. 22-42, 2005.

[2] J. Park, S. Yoon, J. Park, S. Lee, and M. Kim, "Statistic signature based application traffic classification," *J. KICS*, vol. 34, no. 11, pp. 1234-1244, Nov. 2009

[3] B. Park, Y. Won, J. Chung, M. S. Kim, and J. W.-K. Hong, "Fine-grained traffic classification based on functional separation," *Int. J. Network Manag.*, vol. 23, pp. 350-381, Sep. 2013.

[4] IANA port number list, Available: http://www.iana.org/assignments/service-names-prt-numbers/service-names-port-numbers.xml

[5] T. Choi, C. Kim, S. Yoon, J. Park, B. Lee, H. Kim, et al., "Content-aware internet application traffic measurement and analysis," *IEEE/IFIP NOMS 2004*, pp. 511-524, 2004.

[6] N. F. Huang, G. Y. Jai, H. C. Chao, Y. J. Tzang, and H. Y. Chang, "Application traffic classification at the early stage by characterizing application rounds," *Inf. Sci.*, vol. 232, pp. 130-142, May 2013.

[7] W. Huan, H. Lin, H. Li, Y. Zhou, and Y. Wang, "Anomaly detection method based on clustering undersampling and ensemble learning," *2020 IEEE 5th Inf. Technol. and Mechatronics Eng. Conf.* (ITOEC), pp. 980-984, 2020, (https://doi.org/10.1109/ITOEC49072.2020.9141897)

[8] L. Guo, Q. Wu, S. Liu, et al., "Deep learning-based real-time VPN encrypted traffic identification methods," *J. Real-Time Image Process.*, vol. 17, no. 1, pp. 103-114, 2020. (https://doi.org/10.1007/s11554-019-00930-6)

[9] Z. Han, et al., "An effective encrypted traffic classification method based on pruning convolutional neural networks for cloud platform," *2021 2nd Int. CECIT*, pp. 206-211, 2021. (https://doi.org/10.1109/CECIT53797.2021.00043)

[10] A. A. Afuwape, Y. Xu, J. H. Anajemba, and G. Srivastava, "Performance evaluation of secured network traffic classification using a machine learning approach," *Computer Standards & Interfaces*, vol. 78, no. 103545, 2021, ISSN 0920-5489. (https://doi.org/10.1016/j.csi.2021.103545)

[11] L. Breiman, "Random forests machine learning," *View Article PubMed/NCBI Google Scholar*, vol. 45, pp. 5-32, 2001. (https://doi.org/10.1023/A:1010933404324)

[12] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, pp. 1189-1232, 2001.

이 민 성 (Min-Seong Lee)

2020년 : 고려대학교 컴퓨터정보학과 학사

2020년~현재 : 고려대학교 컴퓨터정보학과 석사과정

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

[ORCID:0000-0002-1774-2831]

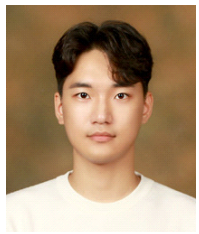
박 지 태 (Jee-Tae Park)

2017년 : 고려대학교 컴퓨터정보학과 학사
2017년~현재 : 고려대학교 컴퓨터정보학과 석박사통합과정
<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석
[ORCID:0000-0002-8515-6164]

백 의 준 (Ui-Jun Baek)

2018년 : 고려대학교 컴퓨터정보학과 학사
2018년~현재 : 고려대학교 컴퓨터정보학과 석박사통합과정
<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석
[ORCID:0000-0002-4358-7839]

최 정 우 (Jung-Woo Choi)



2018년 : 고려대학교 컴퓨터정보학과 학사
2018년~현재 : 고려대학교 컴퓨터정보학과 석사과정
<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석
[ORCID:0000-0002-0492-8311]

김 명 섭 (Myung-Sup Kim)

1998년 : 포항공과대학교 전자계산 학과 학사
2000년 : 포항공과대학교 전자계산 학과 석사
2004년 : 포항공과대학교 전자계산 학과 박사
2006년 : Dept. of ECS, Univ of Toronto Canada
2006년~현재 : 고려대학교 컴퓨터정보학과 교수
<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석, 멀티미디어 네트워크
[ORCID:0000-0002-3809-2057]