# 머신러닝 알고리즘을 적용한 SDN Controller 기초연구

김혜은, 김상철
국민대학교

hyeeun7904@kookmin.ac.kr, sckim7@kookmin.ac.kr

# A Study on the SDN Controller using Machine Learning Algorithm

Kim Hae Eun, Sang-Chul Kim
Kookmin Univ.

요 약

본 논문은 머신러닝 알고리즘을 적용해 트래픽 분류에 관한 연구를 제시하였다. D-ITG 툴을 사용해 Ping, DNS, Telnet, Voice 4 종류의 가상 트래픽을 발생시켰고, Logistic Regression, K-Means, Random Forest 3 개의 머신러닝 알고리즘으로 트래픽 분류를 진행하였다. 본 연구의 결과는 트래픽 분류에 가장 적합한 알고리즘을 찾고 기초 연구를 진행하는데 중요한 역할을 하였다.

## 1 서 론

오늘날 IP 네트워크의 트래픽 flow 분류는 SDN 에 인공지능 기술을 접목해 중요한 연구분야로 부상하고 있다. 포트번호를 기반으로 트래픽을 식별하고 페이로드 검사로 분류하던 기존 방법은 트래픽의 종류가 다양해지고 암호화되면서 제약이 생겼다.

이렇게 네트워크가 복잡해짐에 따라 인공지능 기술을 접목한 트래픽 관리 기술의 연구가 여럿 있는데, 트래픽을 효율적으로 분류하기 위해 기존의 특징선택 알고리즘의 특징 수가 작으면 분류 정확도가 낮아지고 특징 수가 많으면 시간이 오래 걸리게 되는 단점을 보완해 새로운 특징 선택 알고리즘을 구현[1]하거나 실제 네트워크 환경에서 대용량 트래픽을 빠르게 처리하기 위해 앙상블 모델과 딥러닝 모델을 사용해서 먼저 머신러닝 모델을 사용해서 분류하고, 분류가 어려운 데이터는 딥러닝 모델에서 분류하도록 분류기를 구현한 연구[2] 결과도 있다.

이 프로젝트에서는 Logistic Regression, K-means, Random Forest 3 가지 머신러닝 알고리즘을 활용해서 대역폭, Qos 및 다양한 flow level 정보를 기반으로 트래픽을 분류하려고 한다.

데이터셋은 D-ITG (Distributed Internet Traffic Generator) 툴을 이용해 발생시킨 가상 트래픽을 사용한다. 분류하고자 하는 트래픽 종류로 Ping, DNS, Telnet, Voice 를 선택했으며, 4 가지 각 종류별로 트래픽을 발생시켜 각각의 데이터를 수집했다.

## 2 본론

### 2.1 실험환경 소개

가상환경은 컨트롤러 1 개, 스위치 1 개, 호스트 3 개로 구성되어있으며, 각 호스트는 OVS(Open vSwitch) 의 overlay 네트워크를 통해 연결되도록 설계하였다.

가상 트래픽 생성기로는 Mininet 과 D-ITG 가 고려되었는데, Mininet 은 구축된 가상 네트워크 안에서만 트래픽을 생성할 수 있는 툴이기 때문에 네트워크 간 트래픽 시뮬레이션에는 한계가 있었다. 따라서 서로 다른 네트워크 환경 간에 트래픽을 발생시킬 수 있는 D-ITG 툴을 사용해서 실제 네트워크 플로우를 재현하였다.

### 2.2 데이터 소개

수집된 DNS, Ping, Telnet, Voice 데이터는 각자의 역할에 맞게 아래와 같은 특징을 가졌다.

DNS 트래픽은 클라이언트와 DNS 서버 사이의 IP 를 질답하는 트래픽으로, 짧은 핑퐁과 작은 크기의 패킷의 특징을 갖는다.

| 평균 | DNS | Ping | Telnet | Voice |
|---|---|---|---|---|
| Delta Forward Packets | 0.8 | 1.0 | 31 | 0.9 |
| Delta Forward Bytes | 57 | 98 | 2763 | 62 |

Ping 트래픽은 에코 요청을 테스트 하는 트래픽으로 일정한 주기로 반복적으로 발생하며, 기본적으로 패킷 사이즈가 설정되기 때문에 패킷 별 크기가 일정하게 유지되므로 낮은 분산을 띈다.

| 편차 | DNS | Ping | Telnet | Voice |
|---|---|---|---|---|
| Delta Forward Packets | 2.3 | 0.1 | 51.1 | 2.5 |
| Delta Forward Bytes | 162.3 | 8.8 | 4504.9 | 171.5 |

Telnet 트래픽은 원격 시스템의 가상 터미널에 접근하는 트래픽으로 사용자의 명령을 담은 트래픽이다. 따라서 사용자가 명령을 불규칙적으로 보내기 때문에 편차가 굉장히 큼을 확인할 수 있다.

| 편차 | DNS | Ping | Telnet | Voice |
|---|---|---|---|---|
| Forward Instantaneous Packets per Second | 2.3 | 0.1 | 50.9 | 2.5 |
| Forward Average Packets per second | 0.2 | 0.02 | 10.1 | 0.2 |
| Forward Instantaneous Bytes per Second | 162.3 | 9.0 | 4492.6 | 171.5 |
| Forward Average Bytes per second | 20 | 2.3 | 887.7 | 13.9 |

Voice 트래픽은 음성 데이터가 패킷으로 나뉘어 압축된 상태로 전송되기 때문에 패킷 크기가 작게 나타난다. 반면 Reverse 트래픽은 주로 ACK(확인 응답) 패킷이나 네트워크 상태 정보를 포함해, Voice 통신의 품질을 유지하기 위한 제어 정보가 포함되어 있어. 이때 각 패킷에는 상당한 양의 메타데이터가 포함되기 때문에 크기가 비교적 크다.

| 평균 | DNS | Ping | Telnet | Voice |
|---|---|---|---|---|
| Forward Instantaneous Bytes per Second | 57 | 98 | 2743 | 62 |
| Forward Average Bytes per second | 61 | 98 | 2083 | 64 |
| Reverse Instantaneous Bytes per Second | 192 | 98 | 2058 | 7033 |
| Reverse Average Bytes per second | 195 | 98 | 1562 | 6987 |

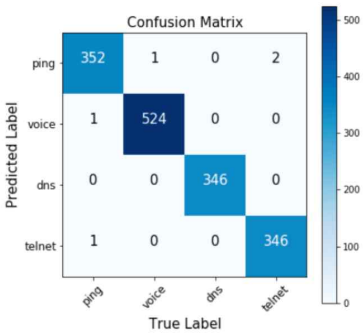
## 2.3 핵심 알고리즘 소개

이 프로젝트에서는 세 가지 머신러닝 알고리즘을 사용한다. 첫 번째로 사용된 Logistic Regression 알고리즘은 지도 학습 모델로, 주로 범주형 타겟을 분류하는 데 사용되며, 데이터가 선형적 분포를 가질수록 성능이 향상된다. 두 번째로 사용된 K-Means 알고리즘은 비지도 학습 모델로, 유클리드 거리를 사용해 클러스터 중심에서 가까운 데이터 포인트를 묶어 클러스터링하는 알고리즘이며, 데이터가 원형에 가까운 클러스터일 수록 잘 동작한다. 마지막으로 사용된 Random Forest 알고리즘은 앙상블 학습 기법으로 여러 결정 트리들의 판단을 종합해 분류하는 모델이다. 데이터가 비선형적이고 복잡한 분포를 가질수록 Random Forest 의 성능이 더 높아진다.

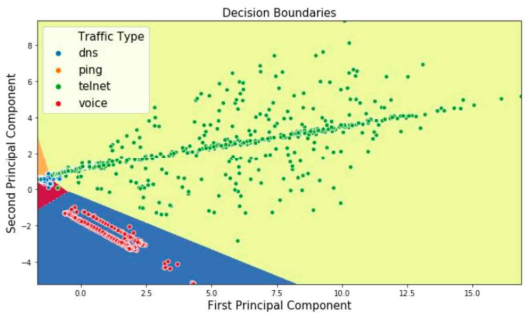
모든 알고리즘에 대해 결측치와 학습에 불필요한 데이터를 삭제해서 데이터를 축소하는 과정을 거쳤고, (총 패킷 수를 나타내는 Forward Packets, Reverse Packets 는 삭제한다.) 학습 데이터와 테스트 데이터의 비율은 70:30 으로 설정했다.

### 2.3.1 Logistic regression 알고리즘

sklearn 라이브러리의 LogisticRegression 클래스를 사용해 모델 학습을 진행한 결과, 99.3% 정확도로 분류했다. confusion matrix 로 시각화하면 다음과 같다 (0=DNS, 1=Ping, 2=Telnet, 3=Voice). True Label 과 Predicted Label 이 거의 일치하는 것을 볼 수 있다.



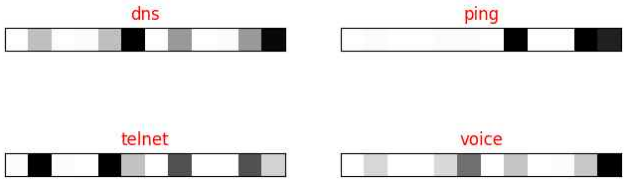

PCA 를 이용해서 2 개의 주요 성분을 추출해 LogisticRegression 모델로 학습시키면 다음과 같은 Decision Boundary PCA 를 반환한다. 2 개의 주요 성분만으로 학습한 LogisticRegression 모델은 84%의 정확도로 데이터를 예측했다.



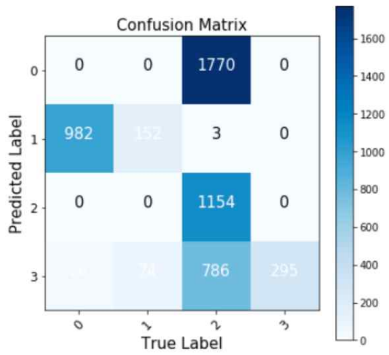

### 2.3.2 K-means clustering 알고리즘

sklearn 라이브러리의 KMeans 클래스를 사용해 모델 학습을 진행한 결과, 30.5% 정확도로 분류했다. 4 개의 클러스터 (DNS, Ping, Telnet, Voice) 로 분류하는데 사용된 12 개의 속성들을 차원으로 두고 12 차원 벡터 공간에서 클러스터 중심이 어디에 있는지 시각화하면

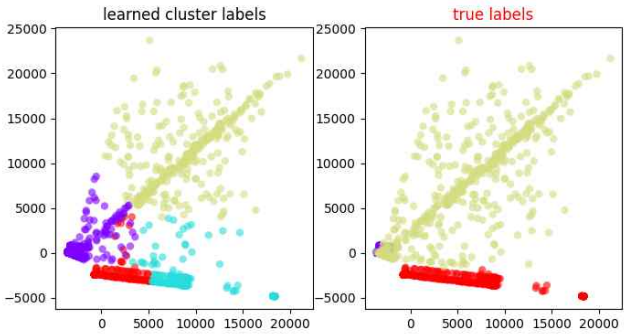
아래 그림과 같이 각 traffic 마다 어떤 feature 가 두드러지는지 확인할 수 있다.




Confusion matrix 로 확인해보면 다음과 같다 (0=DNS, 1=Ping, 2=Telnet, 3=Voice). 여기서 KMeans 모델이 Voice 를 두 개의 레이블로 분류하고 Telnet 은 거의 절반 이상을 잘못 레이블링하고 있는 등 정확하게 분류하지 못하는 모습을 확인할 수 있다.

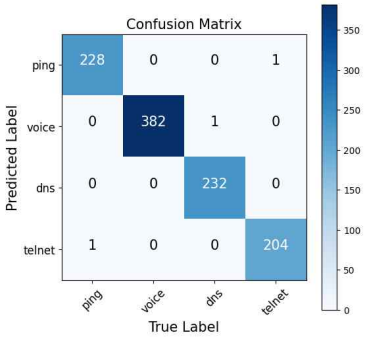



k-means 모델을 사용해서 2 차원 PCA 로 분류하면 다음 사진과 같이 k-means 모델이 Voice(빨강, 하늘)와 Telnet(노란색, 보라색, 하늘색) 을 실제 라벨 값과 달리 여러 개의 레이블로 잘못 나누고 있음을 확인할 수 있다.

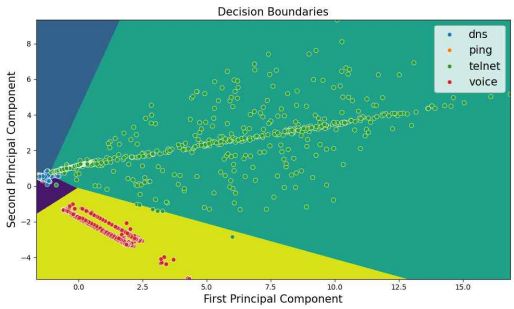



2.3.3 random forest 알고리즘

sklearn 라이브러리의 RandomForestClassifier 클래스를 사용해 트리의 개수를 100 개로 설정하고 모델 학습을 진행한 결과, 99.7% 정확도로 분류했다. confusion matrix 로 시각화하면 다음과 같다 (0=DNS, 1=Ping, 2=Telnet, 3=Voice). True Label 과 Predicted Label 이 거의 일치하는 것을 볼 수 있다.




PCA 를 이용해서 2 개의 주요 성분을 추출해 LogisticRegression 모델로 학습시키면 다음과 같은 Decision Boundary PCA 를 반환한다. 2 개의 주요 성분만으로 학습한 LogisticRegression 모델은 78%의 정확도로 데이터를 예측했다.




## Ⅲ. 결론

데이터는 선형적인 분포를 띄기 때문에 Logistic Regression 이 잘 동작한 것으로 보인다. 반면 Random Forest 는 비선형적 데이터를 잘 분류하지만, 여러 결정 트리의 결과를 반영해 선형적 모델과 비슷한 결과를 낸 것으로 보인다. 또한 원래 Logistic Regression 알고리즘은 이진 변수를 예측하는데 주로 사용되지만, 이 프로젝트에서는 4 개의 변수를 예측하도록 사용되어 발생 확률이 가장 높은 대상을 예측 값으로 선택한다. 이는 Random Forest 의 결정 트리 방식과 유사한데, Random Forest 가 Logistic Regression 보다 더 많은 결정 트리를 갖고 있어서 미미하게 더 높은 정확도로 예측을 한 것으로 보인다. 반면, K-Means 알고리즘은 애초에 라벨을 예측하기 위해 사용하는 알고리즘이 아니고 원형에 가까운 클러스터일 수록 잘 동작하는데, 제공된 데이터는 선형 분포를 띄는 데이터였기에 낮은 예측 정확도를 기록한 것으로 보인다.

.


## ACKNOWLEDGMENT

본 연구는 2022 년 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음"(2022-0-00964)


## 참 고 문 헌


[1] 임환희, 김경태, 이병준, 윤희용. (2019). SDN 환경의 트래픽 분류를 위한 특징 선택 기법. 한국통신학회논문지, 44(1), 106-116, 10.7840/kics.2019.44.1.106

[2] 이민성, 박지태, 백의준, 최정우, 신창의, 김명섭. (2022). 순차적인 데이터 처리를 통한 딥 러닝 기반 트래픽 분류속도 개선. 한국통신학회논문지, 47(12), 2096-2103, 10.7840/kics.2022.47.12.2096