기계학습을 이용한 Office365 서비스 분류 알고리즘 성능 평가

권윤주, 이민성, 최정우, 박지태
고려대학교

{tkwkf12, min0764, choigoya97 pjj5846 }@korea.ac.kr

Performance Evaluation of Office365 Service Traffic Classification Algorithm Using Machine Learning

Yun-Ju Kwon, Min-Seong Lee, Jeong-Woo Choi, Jee-Tae Park
Korea Univ.

요 약

응용 프로그램을 실행할 때 발생하는 네트워크 트래픽들을 수집해 분류하는 것은 수많은 다른 네트워크 활동에 근본적으로 중요한 역할을 한다. 본 논문의 목적은 Decision Tree, Gaussian Naive Bayes 등의 두 가지 지도학습 알고리즘을 사용해 플로우의 통계 정보를 가지고 Office365 응용프로그램을 분류하고 두 알고리즘들의 성능을 서로 비교하는 것이다. SaaS(Software-as-a-Service) 어플리케이션을 실행시켜 모니터링을 한 다음 발생하는 트래픽으로 트래픽 데이터셋을 만든다. 데이터셋을 두 가지 분류모델에 적용해 비교하면 Decision Tree 분류 모델이 어플리케이션을 구별하는 데 Gaussian Naïve Bayes 분류 모델보다 더 효율적일 거라는 결과를 확인할 수 있다.

Ⅰ. 서 론

응용 프로그램을 실행할 때 발생하는 응용 트래픽을 분류하는 일은 오늘날 수많은 네트워크 분야, 즉 보안 모니터링, QoS(Quality of Service), 프로그램 사용자들의 행위탐지 등을 포함한 다양한 네트워크 분야에서 근본적으로 중요한 역할을 한다. 본 논문에서는 많은 네트워크 트래픽 중에서 Office 365 클라우드 서비스의 트래픽을 수집해 사용자가 사용한 office365 서비스가 무엇인지 분류하고 식별한다. Office365 의 대표적인 응용 프로그램인 Word, PowerPoint, Excel 으로 총 3 가지 프로그램을 구분한다. 클라우드 서비스는 사용자에게 인터넷 기반의 컴퓨팅 자원을 빌려주고 사용한 만큼 비용을 청구한다. 기업의 서비스 운영에 비용 절감과 현실적이고 효율적인 자산 활용에 도움을 주기 때문에 오늘날 기업들은 클라우드 서비스를 도입하고 있다.

본 논문은 기계 학습을 사용해 SaaS 클라우드 서비스 중 Office 365 트래픽을 분류하여 기업들이 트래픽 분석하는 일에 현실성과 효율성에 도움을 주고자 한다. Microsoft Network Monitor 를 사용하여 Office 365 응용프로그램의 네트워크 트래픽을 수집한다. 수집한 네트워크 트래픽 cap 파일 형식을 5-tuple 이 일치한 패킷들을 정의한 플로우(flow) 단위로 저장한 FWP(Flow With Packet)파일 형식으로 변환한다. 그런 다음에 플로우 별로 통계정보를 수집하여 데이터 셋으로 만들어 두 Decision Tree 분류 모델과 Gaussian Naïve Bayes 분류 모델에 적용해 정확도와 시간을 비교한다. [1]

[1] 이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(NRF-2018R1D1A1B07045742)과 2020년도 산업통상자원부 및 한국산업기술평가관리원(KEIT) 연구비 지원에 의한 연구임 (No. 20008902, IT비용 최소화를 위한 5채널 탐지기술 기반 SaaS SW Management Platform(SMP) 개발)

이 두 모델은 네트워크 트래픽 플로우를 가지고 사용자들이 3 가지 Office 365 의 응용프로그램 중 어떤 프로그램을 서비스를 이용했는지 식별하는 것을 목표로 한다. Decision Tree 분류 모델이 Gaussian Naïve Bayes 보다 조금 더 높은 정확도를 제공한다는 결과를 보여준다.

본 장을 제외한 나머지 부분에서는 SaaS 클라우드 서비스 트래픽 분류 모델을 단계적으로 살펴보고 두가지 기계학습 분류모델들의 성능을 비교분석하고 마지막에는 실험 결과와 향후 연구방향을 제시한다.

Ⅱ. 본론

본 절에서는 어떻게 SaaS 클라우드 서비스의 트래픽 플로우를 식별하고 분류하는지를 두가지 기계학습 분류모델을 사용하여 단계적으로 설명을 한다.

A. 데이터 셋

실시간으로 발생하는 Office365 의 3 가지 응용 프로그램들(Word, PowerPoint, Excel)의 네트워크 트래픽을 수집한다. 패킷 단위로 수집한 트래픽은 .cap 파일 형식으로 저장을 한 다음에 .cap 파일을 5-tuple(source port, destination port, source ip address, destination ip address, protocol)이 일치한 패킷들의 집합 단위인 플로우(flow)로 [2] 저장한 FWP(Flow With Packet)파일 형식으로 변환을 시킨다. 플로우의 통계정보(패킷의 개수, 패킷 크기, 패킷 발생시간)를 수집을 하고 추출한 통계정보들로 분류 모델 입력에 적합하도록 전처리를 한다. 그런 다음, 수집했던 응용프로그램의 트래픽에 따라 통계정보를 라벨링을 한다. 지도학습 알고리즘은 데이터 세트 중 일부 데이터를 훈련용 데이터로 만들어 모델을 훈련시킨 다음 나머지를 테스트용 데이터로 사용해 예측을 출력하게 되는데 본 논문 실험에서는 전체 데이터셋 중에 테스트 사이즈를 0.3 을 주어 훈련용

데이터로 만들고 나머지는 테스트용 데이터로 만들어 실험한다.

B. 분류 모델

분류 모델로는 Decision tree 와 Gaussian Naïve Bayes 을 사용한다.

Decision Tree 알고리즘은 최소한의 시도와 적은 시간으로 예측을 출력할 수 있기 때문에 트래픽을 분석하는 데 중요한 알고리즘이다.

Gaussian Naïve Bayes 알고리즘은 확률기반으로 하는 예측방법이다. Naïve Bayes 분류 모델은 많은 메모리 량을 요구하지 않고 예측 정확도가 좋기 때문에 다른 분류 모델과 비교하면 성능이 우수하다. [1]

Ⅲ. 실험 및 결과

정확도를 높이는 실험결과를 도출하기위해 응용프로그램을 이진분할과 삼중분할을 하여 정확도를 비교한다. 암호화되지 않은 트래픽 플로우들이 담긴 Common 파일, 암호화된 트래픽 플로우들이 담긴 Encrypted 파일, Common 파일과 Encrypted 파일을 합친 Entire 파일 총 3 가지 데이터셋들을 가지고 각각 두 분류모델을 학습시켜 예측의 정확도를 비교 분석한다.

그림 1 은 Decision Tree 분류모델의 트리 깊이를 1 부터 200 까지 뒀을 때 나오는 정확도와 Gaussian Naïve Bayes 분류 모델을 200 번 돌렸을 때 나오는 정확도의 평균값, 최대값, 최솟값을 비교하여 보여준 표이다.

그림 2 은 Decision Tree 에서 데이터 셋 별로 트리의 최대 깊이에 따라 변하는 응용프로그램 트래픽의 이진분할과 삼진분할의 정확도를 나타낸 그래프이다. Decision Tree 분류모델 Gaussian Naïve Bayes 분류 모델을 200 번 돌렸을 때 나오는 정확도의 평균값, 최대값, 최솟값을 나타낸 것이다.

		Average(%)		Maximum(%)		Minimum(%)	
		DT	NB	DT	NB	DT	NB
Word <-> PPT <-> Excel	Common	62.069	36.827	70.989	43.344	42.32	31.058
	Encrypted	70.018	38.595	78.498	46.416	40.273	31.399
	Entire	65.422	37.75	72.696	30.375	41.979	49.146
Word <-> PPT	Common	74.819	52.871	83.582	41.293	56.218	60.695
	Encrypted	82.287	55.831	90.547	45.771	58.208	64.766
	Entire	77.429	58.486	86.567	45.275	59.701	68.159
PPT <-> Excel	Common	74.518	49.152	84.375	40.625	59.895	58.333
	Encrypted	81.168	58.33	87.5	48.958	60.937	69.27
	Entire	76.69	52.434	84.375	41.666	58.333	64.062
Word <-> Excel	Common	69.908	54.362	77.835	43.298	56.701	61.855
	Encrypted	73.776	51.253	82.989	43.8144	54.639	61.34
	Entire	71.693	54.027	81.443	43.298	60.309	61.34

그림 1 두 분류 모델 정확도와 비교

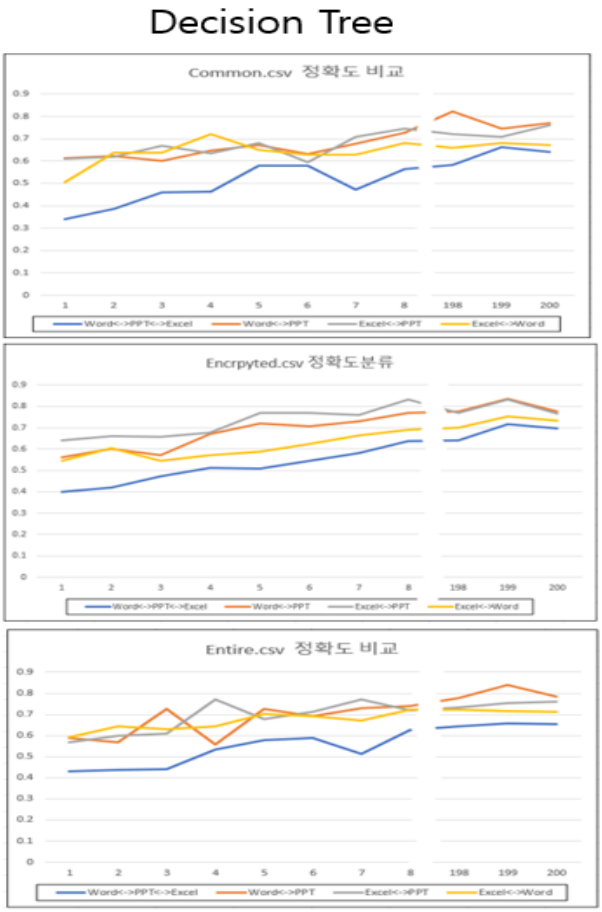

그림 2 트리 깊이에 따른 Decision Tree 정확도 비교

결과는 두 분류모델을 비교해 봤을 때 Decision Tree 분류모델이 Gaussian Naïve Bayes 분류모델보다 훨씬 더 높은 정확도를 보여준다.

Ⅳ. 결론 및 향후 연구 계획

본 논문은 클라우드 서비스 중 Office365 의 트래픽 플로우들로 총 3 가지 응용프로그램(Word, PowerPoint, Excel)으로 분류하기 위한 두가지 기계학습 알고리즘(Gaussian Naïve Bayes, Decision Tree)을 사용해 어떻게 네트워크 트래픽 분류 기술을 적용할지에 대한 방법론을 제시한다.

향후 연구로는 더 효율적인 분류 알고리즘을 찾기 위해 계속 조사하고 비교분석 할 예정이다.

참 고 문 헌

[1] SHAFIQ, Muhammad; YU, Xiangzhan; LAGHARI, Asif Ali. WeChat text messages service flow traffic classification using machine learning technique. In: 2016 6th International Conference on IT Convergence and Security (ICITCS). IEEE, 2016. p. 1-5.

[2] 정구현; 이봉환; 양동민. 패킷 페이로드 내 특정 패턴 탐지 알고리즘들의 성능 분석에 관한 연구. 한국정보통신학회논문지, 2018, 22.5: 794-804.