데이터의 필터링을 위한 머신 러닝 및 앙상블 분류기에서의 신뢰도 측정

이민성, 박지태, 최정우, 백의준, 김명섭

고려대학교

{min0764, pjj5846, choigoya97, pb1069, tmskim}@korea.ac.kr

Measurement of Reliability in Machine Learning and Ensemble Classifiers for Data Filtering

Lee Min-Seong, Park Jee-tae, Choi jeong-woo, Ui-Jun Baek, Kim Myung-Sup

Korea Univ.

요 약

코로나19 사태로 인하여 재택근무 환경, 새로운 응용 트래픽의 등장으로 개인 및 기업에서의 인터넷 사용이 크게 증가하고 있다. 이에 따라 대용량 트래픽을 관리하기 위한 응용 트래픽 분류나 대용량 트래픽을 처리하기 위한 속도에 대한 문제를 해결하기 위한 연구들이 진행되고 있다. 네트워크 트래픽 분류는 네트워크의 효율적인 운용과 관리, 서비스 품질 향상(QoS), 보안 향상의 측면에서 필요한 연구이며 새로운 응용 트래픽들의 등장으로 기존에 사용되던 분류기에 추가적인 응용 트래픽 분류를 할 수 있는 분류기가 생성되어야 한다. 응용 트래픽의 분류를 위하여 머신 러닝 및 앙상블 분류기를 사용한 연구들이 많이 진행되고 있다. 본 논문에서는 처리 속도가 빠른 머신 러닝 및 앙상블 분류기에서 먼저 분류를 진행하고 분류된 데이터를 필터링하기 위한 방법으로 임계값에 대하여 정의하고 정의된 임계값을 기반으로 한 신뢰도를 측정한다. 그리고 응용 트래픽 분야에서 많이 사용되고 있는 분류기들의 신뢰도에 따른 데이터의 처리 개수를 결과로 나타낸다. 실험결과 Random Forest를 사용한 앙상블 분류기에서 가장 높은 신뢰도를 나타내었으며 100%의 신뢰도를 달성하기 위한 임계값은 0.9를 나타내었다.

Ⅰ. 서 론

코로나19 사태로 인하여 네트워크의 활용이 급격하게 변화하고 있다. 재택근무 환경, 새로운 응용 트래픽의 등장 등으로 인하여 개인이나 기업에서의 인터넷 사용이 크게 증가하였다. 특히, 생활 방식이 달라지면서 모바일 및 웹 응용 프로그램들을 사용하는 사람들이 많아지고, 이는 자연스럽게 트래픽의 증가로 이어졌다. 이에 따라 대용량 트래픽을 관리하기 위한 응용 트래픽 분류나 대용량 트래픽을 처리하기 위한 속도에 대한 문제를 해결하기 위한 연구들이 진행되고 있다[1].

네트워크 트래픽 분류는 네트워크의 효율적인 운용과 관리, 서비스 품질 향상(QoS), 보안 향상의 측면에서 필요하다. 네트워크 환경에서 인터넷 사용률이 증가하면서 트래픽의 양이 증가하였고 대용량 트래픽을 처리할 수 있는 방법이 필요하게 되었다. 다양한 응용 프로그램들이 새롭게 등장하면서 기존에 사용되던 분류기에 추가적인 응용 프로그램들을 분류 할 수 있도록 추가해야하는 문제도 발생하고 있다. 분류해야 할 응용 프로그램들이 추가될수록 처리해야하는 트래픽은 많아지게 되고 분류기의 처리 속도에 대한 문제도 해결해야 할 필요가 있다.

응용 트래픽 분류 분야에서 공공 데이터 셋을 사용하여 응용 프로그램들을 분류하는 다양한 연구들이 이루어지고 있으며, 머신 러닝 및 앙상블 분류기를 활용함으로서 높은 분류 정확도를 가진 모델들이 연구되었다[2]. 본 논문에서는 분류기의 처리 속도에 문제점을 두고 데이터를 필터링 할 수 있는 방안에 대하여 고려하였다. 처리 속도가 빠른 머신 러닝 및 앙상블 분류기에서 쉽게 분류 할 수 있는 데이터를 처리하여 분류된 데이터를 필터링하고 어려운 분류 문제들은 다음 분류기로 넘기는 방법을 고안하였다. 이 때 머신 러닝 및 앙상블 분류기에서 처리하는 데이터가 잘 분류되었다는 신뢰성을 확보하기 위하여 임계값을 설정하고 임계값을 기반으로 한 신뢰도를 결정한다. 본 논문에서는 데이터 필터링을 위한 임계값에 대하여 정의하고 트래픽 분류에서 사용되고 있는 머신 러닝 및 앙상블 분류기들을 사용하여 임계값에 대한 신뢰도를 결정하고 데이터 필터링을 할 수 있는 기준으로 사용 할 수 있도록 한다.

본론에서는 실험에 사용하기 위한 데이터 셋에 대한 내용과 실험에 사용되는 머신 러닝 및 앙상블 분류기에 대한 설명을 기술한다. 데이터 필터링을 위하여 제안된 임계값을 설정과 임계값에 의한 신뢰도를 결정하는 방법에 대하여 기술하고 실험 결과를 보여준다. 마지막으로 결론 및 향후 연구로 본 논문을 마친다.

Ⅱ. 본 론

본 장에서는 실험에 사용하기 위한 데이터 셋과 머신 러닝 및 앙상블 분류기에 대하여 간단하게 설명한다. 임계값에 대한 정의와 함께 사용된 분류기들에 따라 정의된 임계값과 임계값에 따른 신뢰도를 측정하고 처리할 수 있는 데이터의 개수를 비교한다.

표 1. 수집한 응용 프로그램

응용 프로그램 유형	응용 프로그램
SNS	Instagram
	Facebook
	Naver_Blog
SHOPPING	11st
	Coupang
	Gmarket
MAIL	Gmail
	Naver_mail
	Nate_mail
STREAMING	Naver_TV

이 논문은 2020년도 산업통상자원부 및 한국산업기술평가관리원(KEIT) 연구비 지원에 의한 연구 (No. 20008902, IT비용 최소화를 위한 5채널 탐지기술 기반 SaaS SW Management Platform(SMP) 개발) 이고, 2021년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과 (2021RIS-004)이다.

A. 데이터 셋

본 절에서는 실험에서 사용된 데이터 셋에 대하여 설명한다. 데이터 셋은 일상생활에서 사용하고 있는 10가지종류의 응용 프로그램들의 500개의 플로우를 수집하여 총 5000개의 플로우를 사용하였다. 5000개의 데이터 중에서 70%는 학습 데이터로, 30%는 테스트 데이터로 사용하였다. 수집한 응용 프로그램의 종류는 표1에 나타나있다.

B. 머신 러닝 및 앙상블 분류기

본 절에서는 실험에 사용한 머신러닝 및 앙상블 분류기는 응용 트래픽 분류 연구에서 사용되고 있는 분류기이다[2]. 사용된 분류기의 종류는 표2에 나타나있다.

표 2. 머신 러닝 및 앙상블 분류기

모델	분류기
Machine Learning	Decision Tree
	K-NN
	Logistic Regression
Ensemble Model	AdaBoost
	Random Forest

C. 임계값과 신뢰도

본 절에서는 임계값을 설정하는 방법과 설정된 임계값을 통해 측정되는 신뢰도에 대하여 설명하고 임계값과 신뢰도 사이의 상관관계에 대하여 기술한다.

신뢰도를 결정하기 위한 임계값은 분류기가 클래스를 분류 시 분류를 확정하는 값을 기준으로 임계값을 결정한다. 분류기가 데이터를 입력으로 받고 클래스를 분류할 때 해당 데이터를 클래스별로 판단하는 분류 정확도를 나타내게 되고, 가장 높은 정확도를 가진 클래스로 판단한다. 임계값은 정확도에 기준을 두어 낮은 정확도를 바탕으로 분류된 데이터는 해당 분류기에서 판단하기 어려운 데이터라고 판단하게 된다.

신뢰도는 임계값을 설정한 후에 분류기가 정답을 정확하게 분류하였는지 확인 할 수 있는 지표이다. 임계값보다 큰 정확도로 클래스를 판단한 데이터들 중 정확하게 분류한 클래스의 비율로 신뢰도를 측정한다. 예를 들어 0.7의 임계값으로 설정하였을 때 분류기가 100개의 데이터를 70%이상으로 분류하였을 때, 정확하게 분류된 데이터가 94개의 데이터가 되면 0.7의 임계값을 둔 분류기의 신뢰도는 94%이다.

D. 실험 결과

본 절에서는 3가지 머신 러닝 분류기 및 3가지 앙상블 분류기 총 5가지 분류기를 사용하여 10가지 응용 트래픽의 분류를 실험한 결과에 대하여 기술한다.

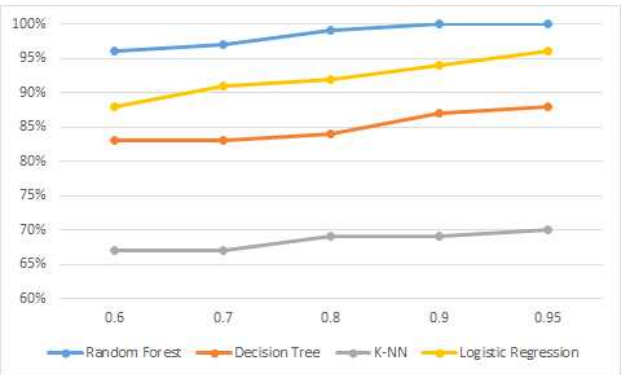

그림 1. 임계값별 측정된 신뢰도

표 3. 임계값 별 처리된 데이터 개수 및 신뢰도

분류기	임계값	0.6	0.7	0.8	0.9	0.95
RF	TP	597	558	420	292	32
	신뢰도	96%	97%	99%	100%	**100%**
DT	TP	918	906	810	745	645
	신뢰도	83%	83%	84%	87%	88%
K-NN	TP	**1050**	**1050**	**1050**	**1050**	**1050**
	신뢰도	70%	70%	70%	70%	70%
LR	TP	384	322	242	137	72
	신뢰도	88%	91%	92%	94%	96%

실험 결과는 그림1과 표3과 같다. 모든 분류기에서 임계값이 높을수록 높은 신뢰도를 보여주고 있다. 분류 정확도를 표기하지 않았으나 Random Forest분류기는 가장 높은 분류 정확도를 보여주고 있고 100%의 신뢰도도 달성하였다. Decision Tree는 많은 데이터를 분류 할 수 있지만 신뢰도가 88%이다. K-NN은 분류기의 특성 상 클러스터를 통해 클래스를 분류하기 때문에 모든 데이터에 대해 분류 정확도가 1이 나오게 된다. 따라서 많은 데이터를 분류하였지만 의미를 두기 어렵다. 마지막으로 그림과 표에 나타나지 않은 AdaBoost의 경우 가장 낮은 분류 정확도를 나타내었으며, 10가지 클래스를 분류하는데 큰 차이가 없는 분류 정확도를 보이고 있어 정확하게 분류하지 못한 분류기라고 볼 수 있다.

임계값과 신뢰도의 상관관계는 임계값이 높아질수록 신뢰도는 높아지게 된다. 하지만 정확하게 분류된 데이터는 줄어들게 되고, 이는 처리된 데이터의 개수가 적어지고 필터링 이후에 남겨진 데이터가 많아지는 것이다. 본 논문의 실험의 목적은 처리 속도를 개선하기 위하여 데이터를 필터링하는 것이다. 따라서 실험 결과를 바탕으로 적절한 신뢰도를 사용하여 처리 속도가 빠른 머신 러닝 및 앙상블 분류기에서 최대한 많은 양의 데이터를 처리하고, 어려운 분류 문제를 잘 다룰 수 있는 분류기에서 남은 데이터를 처리 할 수 있도록 하는 것이다.

Ⅲ. 결론 및 향후 연구

본 논문은 응용 트래픽 분류 문제에서 트래픽의 처리 속도를 고려하고자 데이터 필터링을 적용 할 수 있는 방안에 대해 실험한 논문이다. 처리 속도가 빠른 머신 러닝 및 앙상블 모델에서 데이터를 필터링하기 위하여 분류기에서 나타내는 분류 정확도를 기준으로 임계값을 설정하고 설정된 임계값을 통하여 신뢰도를 측정하였다. 총 5가지의 머신 러닝 및 앙상블 분류기를 사용하여 신뢰도를 측정하였다. 이를 통해 신뢰도를 기반으로 사용하는 분류기의 임계값을 설정하여 데이터를 필터링 할 수 있다.

향후 연구로는 실험한 머신 러닝 및 앙상블 분류기에서 데이터를 필터링 한 후 남은 데이터를 딥 러닝 모델에서 처리하여 전체 처리 속도 비교와 모델의 정확도를 비교할 계획이다.

참 고 문 헌

[1] S. Rezaei and X. Liu, "Deep Learning for Encrypted Traffic Classification: An Overview," in *IEEE Communications Magazine*, vol. 57, no. 5, pp. 76-81, May 2019, doi: 10.1109/MCOM.2019.1800819.

[2] A.Afuwape, Y.Xu, J.Anajemba, G.Srivastava, "Performance evaluation of secured network traffic classification using a machine learning approach", Computer Standards & Interfaces, Volume 78, 2021, 103545, ISSN 0920-5489.