SDN 환경에서 효율적인 트래픽 분류를 위한 Feature Selection 기반 Multi Classification 기법 연구

김기태○ 홍충선*
경희대학교 컴퓨터공학과
{glideslope○, cshong* }@khu.ac.kr

Machine Learning and Feature Based Traffic Classification in Software Defined Network Environment

Kitae Kim○ Choongseon Hong*
Department of Computer Science and Engineering, Kyung Hee University

요 약

IT기술의 발전으로 복잡해진 네트워크 인프라를 효율적으로 관리하기 위해서 Sofrware Defined Network(SDN)의 등장하였으며 이에 따라서 네트워크 관리자는 손쉽게 네트워크를 관리하고 제어할 수 있게 되었다. 관리 기법 중 트래픽 분류는 폭증하는 인터넷 트래픽들의 Quality of Service(QoS)를 제공하기 위한 핵심적인 기술이다. 따라서 본 논문에서는 Feature Selection를 통해 트래픽을 보다 효율적으로 구분할 수 있는 특징을 선택 후 이 결과에 따른 Deep Packet Inspection(DPI)와 Machine Learning 기법을 선택 및 적용을 통한 효율적인 트래픽 분류 기법을 제안한다.

1. 서 론

오늘날 IT기술의 발전에 따라서 스마트폰, 스마트 워치와, Iot 디바이스 같은 다양한 디바이스들의 등장에 따라서 다양한 서비스와 컨텐츠들이 등장하고 있다. 이에 따라 네트워크 상의 데이터들은 폭발적으로 증가하고 있으며 시스코에 따르면 이러한 전 세계 인터넷 트래픽이 2019년에는 2014년 대비 3배 이상인 168EB(엑사바이트)에 다다를 것으로 예측되고 있다. 따라서 폭증하는 트래픽을 효율적으로 관리하며 사용자의 요구를 만족시킬 수 있는 네트워크 관리 기술이 주목받고 있으며 Software Defined Network(SDN)는 앞서 말한 조건을 충족시킬 수 있는 차세대 네트워킹 기술이다. SDN의 가장 큰 특징은 데이터 전송을 담당하는 Data Plane과 Control Layer을 분리시켜 네트워크 관리자가 프로그래밍을 통해 네트워크 관리가 가능하다는 점이다[1]. 이를 이용해서 앞서 언급된 다양한 디바이스에서 오고 가는 예측 불가능한 트래픽들에 대한 효과적인 관리가 가능하고 Quality of Service(QoS)에 따른 최적의 경로를 보장할 수 있다. 이러한 최적의 경로를 탐색하기에 앞서 각 트래픽의 QoS를 분류 할 수 있는 트래픽 분류 기법이 중요하다.

현재 제안된 트래픽 분류기법에는 포트번호를 이용한 기반 트래픽 분류와 패킷의 페이로드를 직접 분석하여 패킷을 분류하는 Deep Packet Inspection(DPI) 기법이 제안되어 왔고 최근 연구에서는 앞서 언급한 두 가지 방법의 한계를 극복하고 폭증하는 다양한 트래픽을 분류하기 위해 Machine Learning(ML) 기법이 제안되고 있다[2]. [3]에서는 트래픽의 정확한 분류를 위해 DPI기법과 ML기법을 혼용해서 사용한다. 하지만 인입되는 패킷에 대해 두 기법중 하나를 선택하기 위한 ML의 분류 결과 Reliability Threshold 값을 계산해줘야 하기 때문에 초기 몇 개의 패킷에 대해서 두 가지 기법을 모두 실행하기 때문에 데이터 트래픽이 많아지게 되면 컨트롤러의 자원을 사용함에 있어서 오버헤드가 있으며 주어진 데이터 셋을 그대로 ML알고리즘에 적용시킴에 따라 높은 정확도를 기대하기 힘들다. 따라서 본 논문에서는 두 기법을 이용하되 어떤 기법을 이용하는지 결정하는데 있어서 먼저 Feature Selection을 통해 분류의 기준이 될 수 있는 Feature만을 추출해 낸 후 ML알고리즘을 적용시켜 ML 알고리즘의 계산 오버헤드를 줄이는 동시에 정확도 높은 트래픽 분류기법을 제안한다.

2. 관련연구

2.1 특징선택(Feature Selection)

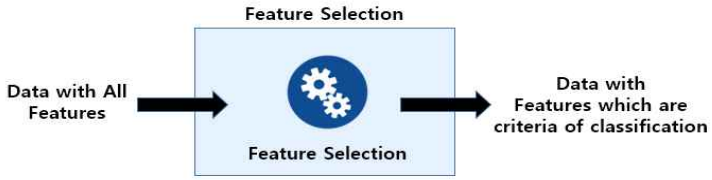

그림 1 Feature Selection

이 논문은 2016년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (B0190-16-2013, 유무선 통합 네트워크에서 접속 방식에 독립적인 차세대 네트워킹 기술 개발)

*Dr. CS Hong is the corresponding author

특징선택(Feature Selection)은 기계학습에서 분류정확도를 향상시키기 위해 원본데이터가 주어졌을 때, 가장 좋은 분류 성능을 보여줄 수 있는 데이터의 부분집합을 원본 데이터에서 찾아내는 방법이다[4]. 이러한 과정을 통해서 첫 번 째로 분류를 위한 정보만을 추출해 내기 때문에 해당 데이터로 분류 알고리즘에 적용 시켰을 시에 분류 정확도를 향상 시킬 수 있으며 두 번째로 분류에 필요하지 않은 데이터를 없앰으로서 원본데이터를 분류를 했을 때 보다 더욱 빠르게 계산을 끝낼 수 있다. 따라서 본 논문에서는 이 특징선택을 통해 각 트래픽을 구분 지을 수 있는 특징을 추출해 낸 후 머신러닝 알고리즘을 적용시켜 빠르고 정확한 분류를 할 수 있는 방법을 제안하며 특징선택으로 도출된 특징에 해당하지 않은 한 패킷에 대해서는 DPI기법을 적용해 분류정확도를 높였다.

2.2 결정트리(Decision Tree)

결정트리는 계산비용이 적으며 대규모 데이터를 분류하는데 적합한 머신러닝 알고리즘으로 분류 기술 중 가장 일반적으로 사용되는 방법이다.

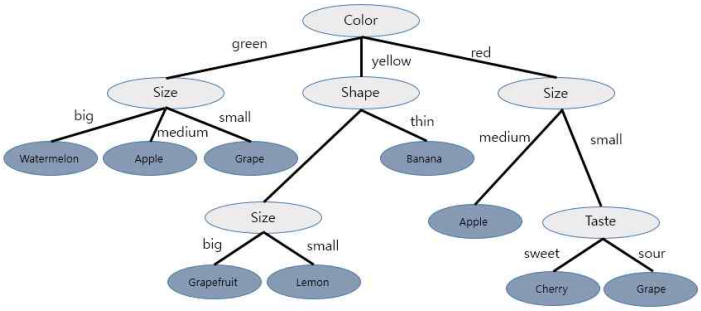

그림2 결정트리(Decision Tree) 구조

그림2는 결정트리의 구조로서 회색 원 안에는 데이터를 구분 지을 수 있는 기준이 들어가게 된다. 본 논문에서는 앞서 2.1절에서 언급되었던 Feature Selection을 이용해 트래픽의 특징을 도출 후 이 결과가 트래픽을 분류하는 기준으로서 회색 원 안에 Decision Attribute로 들어가게 되고 이 기준에 따라 이후에 인입되는 트래픽이 분류가 된다.

3. 제안사항

3.1 Data Feature Selection

본 장은 통계 기반 Feature Selection 기법을 통해 주어진 데이터를 가장 잘 분류할 수 있는 데이터의 Feature를 추출해 낸다. 다음은 주어진 데이터를 통해 통계를 바탕으로 Feature를 추출하는 과정을 나타내는 과정이다.

변수	내용
Q	Matrix : Each row represents Characteristics. Each column represents Samples.
i	Class
j	Feature
M_{ij}	Mean of Each characteristics
CV	Coefficient of Variation
IN_CV_{ij}	CV of each characteristic
OUT_CV_j	CV among all classes by the M_{ij}
CI_j	Definition of coefficient of importance

표 1 Feature Selection 알고리즘에 사용 된 변수

알고리즘1. Feature Selection

1: Discretization of each column of Q;
2: Calculation of coefficient of variation IN_CV_{ij}
3: Computation of coefficient of variation OUT_CV_j
4: Computation $CI_{ij} = \frac{OUT\_CV_j}{IN\_CV_{ij}}$
5: Calculation of mean CI_j = for fiexd j
6: Ordering of CI_j

위 알고리즘에서는 주어진 데이터의 분산의 정도를 나타내는 Coefficient of variation(CV)를 사용하여 Feature Selection을 진행한다. CV가 크면 데이터들은 분산의 정도가 큼을 의미하고 작으면 데이터들은 분산의 정도가 작음을 의미한다. 위 알고리즘에서는 각각의 데이터가 가지고 있는 Feature 값의 CV인 OUT_CV_j와 각각의 Feature의 CV인 IN_CV_{ij}를 구한 후 중요도 계수로 정의된 CI_j를 값을 구하여 각각의 Feature들을 중요도 순서대로 추출할 수 있다. [5].

3.2 Feature Selection Based Decision Tree

Feature Selection 과정을 통해 추출된 Feature를 바탕으로 Machine Learning 알고리즘인 Decision Tree를 생성한다. 앞서 언급된 알고리즘을 통해 Feature는 여러 개가 생성될 수 있으며 생성 된 Feature를 가지고 Decision Tree를 형성 할 때 Feature를 배치하는 순서에 따라 분류 정확도가 상이 할 것이다. 따라서 본 논문에서는 추출된 Feature들 중에서도 분할하기에 가장 좋은 속성의 정도를 측정하는 섀넌 엔트로피를 통해 효율적으로 Decision Tree를 생성한다.

변수	내용
x	Feature
$H(x)$	x 속성의 엔트로피
$p(x)$	x가 나올 확률
$l(x)$	$\log_2 p(x)$
$H(S)$	전체 엔트로피
information gain	정보 이득

표2 새넌 엔트로피 계산에 사용되는 변수

$$H(X) = -\sum p(x)l(x)$$

$$\text{information gain} = H(S) - \sum p(x)l(x)$$

추출된 각 Feature에 대해 Information Gain을 측정한다. Information Gain이란 어떤 Information이 의사결정을 하는데 미치는 영향의 정도를 측정할 수 있는 값이며 이 값이 큰 Feature는 의사결정에 큰 영향을 미치는 Feature이며 이를 바탕으로 트리를 구성하는데 상위에 노드에 배치하는 것이 분류를 함에 있어서 더욱더 정확한 결과를 낼 수 있음을 의미한다. 측정된 Information Gain을 바탕으로 트리를 구성해 Decision Tree의 분류 정확도를 향상시킬 수 있다.

4. 성능평가

4.1 모듈의 구조

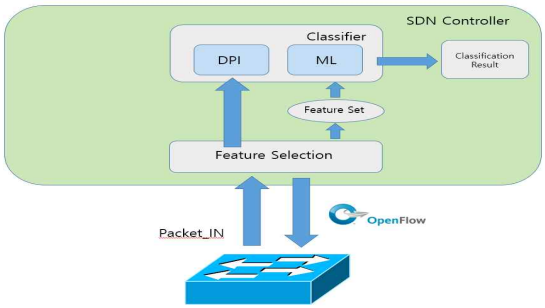

그림 3 시스템 모델

위 그림 3은 최종 모듈에 대한 시스템 모델이다. 스위치가 SDN Controller로 Packet_IN 메시지를 보내고 해당 패킷이 Feature Set에 있는 Feature를 통해 분류가 가능하다고 판단할 경우 ML알고리즘을 통해 분류가 되며 그렇지 않는 경우에는 DPI를 통한 분류가 진행된다.

4.2 평가

본 장에서는 제안했던 모듈에 대한 성능을 검증한다. Youtube, Youtube Live, Skype, TED, Facebook 각각 패킷으로 이루어진 데이터셋[6]을 이용해 Feature Selection을 통해 Feature를 추출하고 이를 바탕으로 Decision Tree를 형성 후 각 패킷을 제안된 모듈에 적용 후 어플리케이션 별 패킷 총 개수 대비 정확하게 분류된 패킷의 비율을 측정한다. 그 결과 본 논문에서 제안하는 방법을 사용하였을 때와 DPI 혹은 머신러닝 알고리즘을 단독으로 사용하였을 때를 비교해본 결과 Machine Learning을 단독으로 사용하는 것 보다 높은 분류 정확도를 보였다.

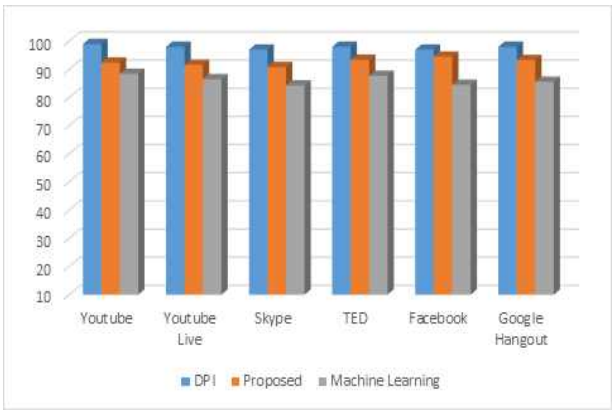

그림 4 패킷 분류 결과

5. 결론 및 향후 연구

본 논문에서는 Feature Selection 기반 Decision Tree와 DPI 모듈을 이용해 정확도 높은 패킷 분류 기법을 제안하였다. Feature Selection을 통해 데이터가 가장 잘 분류될 수 있는 Feature를 찾아낸 후 Machine Learning 혹은 DPI 알고리즘을 적용을 통해 Machine Learning 단독으로 적용했을 때 보다 높은 분류 정확도를 얻을 수 있었다.

향후 연구에서는 DPI와 Machine Learning 기법으로 분류되지 않는 Unknown Packet에 대한 처리, 분류 된 패킷을 바탕으로 SDN 환경에서 빠른 전송경로를 보장할 수 있는 방법에 대한 알고리즘을 연구 할 예정이다.

참 고 문 헌

[1] “소프트웨어 정의 네트워크(SDN)”, http://terms.naver.com entry/.nhn?docId=3580924&cid=59088&categoryId=59096

[2] Pedro Amarla et al., “Machine Learning in Software Defined Networks : Data Collection and Traffic Classification, 2016 IEEE 24th International Conference on Network Protocols(ICNP), 8-11 Nov. 2016

[3] Yuncun LI et al., “MultiClassifier: A Combination of DPI and ML for application-layer classification in SDN, 2014 2nd International Conference on Systems and Informations(ICSAI 2014), 15-17 Nov. 2014

[4] “Feature Selection”, http://mlab.sogang.ac.kr/?mid=researc h_subj_fs

[5] Yu-ning Dong et al “Fine Grained Classification of Internet Multimedia Traffics”, Advanced Communication Technology(ICACT), 2017 19th International Conference on, 19-22 Feb. 2017

[6] Satadal Sengupta, Harshit Gupta, Niloy Ganguly, Bivas Mitra, Pradipta De, Sandip Chakraborty, CRAWDAD dataset iitkgp/apptraffic (v. 2015-11-26), traceset: apptraffictraces, downloaded from http://crawdad.org/iitkgp /apptraffic/20151126/apptraffictraces, http://doi.org/10.1578 3/C77S3W, Nov 2015.