FloFex : 기계학습 기반 응용 트래픽 분류를 위한 트래픽 학습 특징 추출 시스템

박지태, 심규석, 이성호, 김명섭

고려대학교

{pjj5846, kusuk007, gaek5, tmskim} @korea.ac.kr

FloFex : Traffic Learning Feature Extraction System for Accurate Application Traffic Classification based on Machine Learning

Jee-Tae Park, Kyu-Seok Shim, Sung-Ho Lee, Myung-Sup Kim

Korea Univ.

요 약

최근 네트워크 환경의 비약적인 발전과 보급으로 인해 인터넷 트래픽이 급증하고 있다. 이에 응용 별 트래픽 분류의 중요성이 커지고 있으며 이에 대한 다양한 방법들이 연구 되고 있다. 최근에 대두되고 있는 방법 중 하나는 기계학습 기반 응용 트래픽 분류 방법이다. 기계 학습 기반으로 분류 하는 방법에서 가장 중요한 것은 사용 할 학습 특징(Learning Feature)을 적절하게 선정하는 것이다. 어느 학습 특징을 사용하느냐에 따라서 응용 트래픽 분류 결과가 판이하게 달라진다. 이에 본 논문은 기계학습 기반의 응용 트래픽 분류 시스템을 간단히 소개한다. 이 후 소개한 시스템에 사용할 응용 트래픽 학습 특징 추출 시스템을 제안한다. 제안한 방법을 통해 기계 학습에 사용할 응용 트래픽의 학습 특징이 정확하게 나오는지 확인 후 본 논문을 마친다.

Ⅰ. 서 론

최근 네트워크 환경의 발전과 사용 이용자의 증가로 인해 다양한 응용이 발생한다. 이러한 다양한 응용은 각기 다른 패턴의 트래픽을 발생시키고 이러한 패턴은 날이 갈수록 복잡해지고 있다. 따라서 이러한 다양한 패턴의 트래픽을 정확하게 분류하는 방법의 중요성이 커지고 있다.

이에 따라 응용 트래픽을 분류하는 방법은 다양하게 연구되고 있고, 대부분의 방법이 페이로드 시그니쳐 기반의 분류 방법이다[1,2]. 하지만 이러한 시그니쳐 기반의 분류 방법은 기존의 패턴 이외 새로운 패턴이 나오면 대처하기 힘들다는 한계를 가지고 있다. 따라서 최근에는 이러한 한계점을 보완하기 위하여 기계학습 기반으로 분류하는 방법이 대두되고 있다. 기계학습 기반으로 분류하는 방법의 장점은 기존의 시그니쳐 기반의 한계점을 해결 할 수 있고, 오늘날 여러 기업에서 OpenAI 툴을 제공해 주기 때문에 누구든지 자유롭게 사용할 수 있다는 점이다. 본 논문에서는 여러 OpenAI툴 중 Google에서 제공하는 Tensorflow를 사용한다[1].

기계학습 기반의 응용 트래픽 분류 방법은 여러 가지 장점이 있지만 반면에 주의할 사항도 있다. 먼저 기계학습에 사용할 트래픽의 학습 특징 잘 선정해야 한다. 잘못된 학습 특징을 적용하여 실험을 하면 오히려 기존의 응용 트래픽 분류 방법 보다 정확도 면에서 떨어진다. 게다가 기계학습에 사용할 학습 특징이 적으면 가장 최적인 학습 특징을 찾기 힘들다. 최선의 결과를 위해서는 학습 특징을 다양하게 추출해서 여러 번의 실험을 통해 가장 최적인 학습 특징을 찾는 것이 중요하다[1]. 하지만 기존의 응용 트래픽 학습 특징을 추출하는 방법은 추출되는 학습 특징의 개수가 제한적이므로 기계학습 기반의 정확한 응용 트래픽 분류 모델을 만들기에는 적합하지 않다.

따라서 본 논문에서는 기계학습 기반의 응용 트래픽 분류 시스템을 소개하고, 기계학습에 사용할 학습 특징을 다양하고 효율적으로 추출 할 수 있는 시스템을 제안 한다.

Ⅱ. 본론

본 논문에서 제안하는 응용 트래픽 분류 방안은 크게 세 가지로 구성되어 있다.

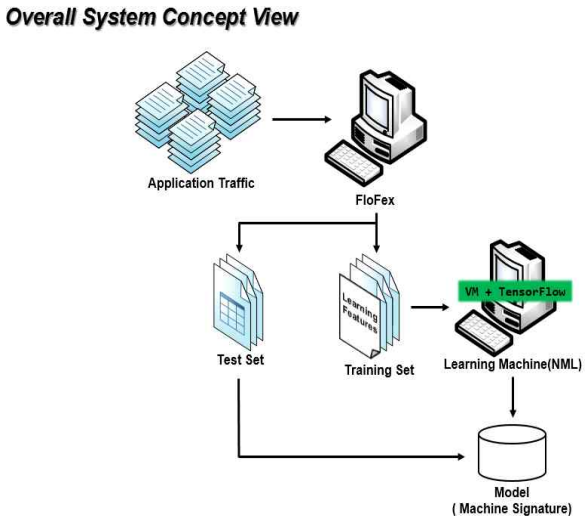

그림 1. 기계학습 기반 응용 트래픽 분류 시스템 구조

먼저 분류할 다양한 종류의 응용 트래픽을 수집한다. 그 후 본 논문에서

이 논문은 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 과학기술인문융합연구사업(No.NRF-2016M3C1B6929228)과 2015년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No.2015R1D1A3A01018057).

제안하는 학습 특징 추출 시스템으로 다양한 학습 특징을 추출하고, 이를 Training Set과 Test Set으로 나눈다. 다음으로 Training Set으로 학습하고 이를 바탕으로 Test Set의 응용 트래픽을 분류 한다.

본 논문에서 제안하는 응용 트래픽 학습 특징 추출 시스템 구조는 그림 2와 같이 구성되어있다. 시스템은 크게 두 가지로 나눌 수 있다. 첫 번째 단계는 Pre-Processing 단계이고, 두 번째 단계는 응용 트래픽의 학습 특징을 추출하는 단계이다.

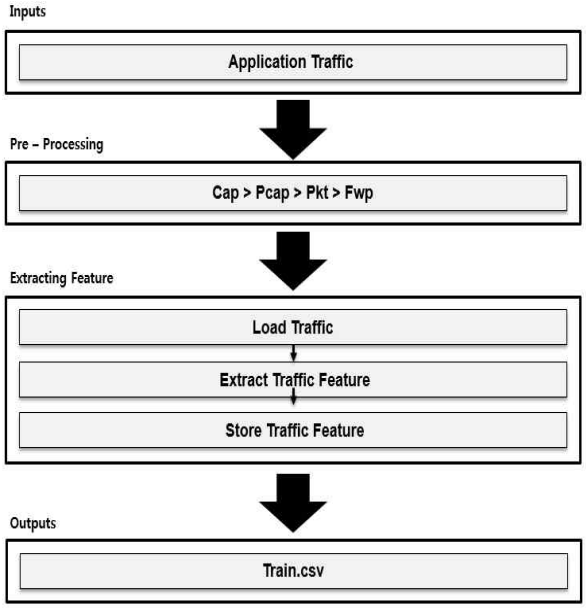

그림 2. FloFex 시스템 구조

먼저 inputs으로 캡쳐된 응용 트래픽인 (Cap File)을 넣는다. 이 후 inputs에 Pre-Processing과정을 적용 시킨 후 트래픽을 두 번째 단계에서 불러온다. 다음으로 불러온 트래픽을 바탕으로 하나의 플로우씩 읽으면서 각 플로우의 학습 특징들을 추출한다. 이 과정에서 Sub Feature는 바로 추출되고 Main Feature은 각각의 값들을 계산하여 추출된다. 마지막으로 추출된 응용 트래픽의 학습 특징들을 output에 저장한다.

Ⅲ. 실험

본 논문에서 제안하는 응용 트래픽 학습 특징 추출 시스템에서 추출 할 수 있는 학습 특징의 구조는 그림 3과 같다. 본 논문에서 제안한 방법으로 추출되는 학습 특징은 약 300개가 있고 각각의 학습 특징은 하나의 플로우 별로 추출된다. 제안한 방법을 통해 학습 특징을 추출한 결과, 추출된 학습 특징들은 크게 Main Feature와 Sub Feature로 나눌 수 있다.

Main Feature는 크게 시간대 별로 6가지로 나눌 수 있는데 이는 전체 시간과 1~5초 이내의 Packet Size와 Inter Arrival Time 학습 특징의 값이다. 각 시간대별로 나오는 Packet Size와 Inter Arrival Time은 또다시 크게 total, forward, backward로 나눌 수 있다. 여기서 패킷(total, forward, backward)은 그림 3과 같이 각각 8 가지의 통계적인 값 (max, min, median, sum, mean, variance 1^{st}quartile,3^{rd}quartile) 을 가진다. max는 최대값, min은 최소값, median은 중간값을 나타낸다. 그리고 sum, mean, variance는 각각 합, 평균, 분산을 나타내고 1^{st}quartile, 3^{rd}quartile는 각각 해당하는 값들의 25%, 75%에 해당하는 값이다.

Sub Feature는 Main Feature와 다르게 응용 트래픽 값의 통계적인 값이 아니라 응용 트래픽에서 쉽게 확인 할 수 있는 정보들로 구성되어 있다. Sub Feature은 그림 3과 같이 크게 네 가지로 나눌 수 있다. 먼저 Packet Count는 각 플로우에 있는 패킷의 개수이고 이는 total, forward, backward로 나눌 수 있다. 두 번째로 Flow Size, Duration, Protocol은 각 플로우의 크기(byte 수), 지속시간, 프로토콜에 해당한다. 세 번째로 Source Port, Source Address와 Destination Port, Destination Address는 각각의 플로우에 해당하는 Source, Destination의 포트번호와 주소에 해당한다. 마지막으로 PPS (Packet Per Second)는 1초 동안 보내는 packet의 수라고 정의한다. PPS도 Packet Count와 마찬가지로 total packet, forward, packet, backward packet에 따라 PPS, FPPS BPPS 로 나누어진다.

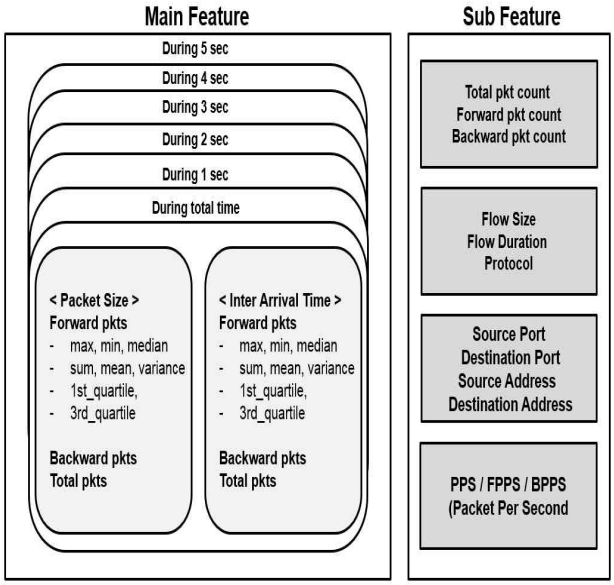

그림 3. FloFex에서 추출된 학습 특징 리스트 구조

Ⅳ. 결론 및 향후과제

본 논문은 정확한 응용 트래픽 분류를 위한 여러 가지 응용 트래픽의 올바른 학습 특징을 다양하고 효율적으로 추출 할 수 있는 방법을 제안하였다. 실험을 통해 input으로 넣은 각 응용 트래픽의 값 (packet count, packet size 등)들과 위 방법으로 얻은 다양한 학습 특징들을 비교해본 결과 정확하게 일치하는 것을 알 수 있었다. 이를 통해 본 논문에서 제안한 시스템을 통해 기계 학습에 사용할 응용 트래픽의 학습 특징을 효율적으로 구할 수 있는 것을 확인 할 수 있었다.

향후 보다 효율적이고 정교한 알고리즘을 사용하여 FloFex 시스템 구조를 개선시킬 것이다. 그리고 여기서 추출된 학습 특징에 Tensorflow의 Softmax Regression을 적용하여 기계학습 기반의 응용 트래픽 분류 모델을 만들 계획이다,

참 고 문 헌

[1]이성호, 심규석, 구영훈, 김명섭, “ Tensorflow 기계학습 도구를 이용한 응용 트래픽 분류”, Nov, 18, 2016, pp224-225.

[2] Sung-Ho Yoon, Kyu-Seok Shim, Su-Kang Lee, and MyungSup Kim , "Framework for Multi-Level Application Traffic Identification," Proc. of the Asia-Pacific Network Operations and Management Symposium (APNOMS) 2015, Busan, Korea, Aug. 19-21, 2015, pp.424-427