점진적 특징 추가를 통한 네트워크 트래픽 이상 탐지 성능 향상에 관한 연구

최다영°, 이주홍*, 박형곤°*
이화여자대학교 전자전기공학전공°, 이화여자대학교 전자전기공학과 스마트팩토리융합전공*
{dddaynz, joohong.rheey}@ewhain.net, hyunggon.park@ewha.ac.kr

A Study on Performance Improvement of Network Traffic Anomaly Detection via Progressive Feature Addition

Dayoung Choi°, Joohong Rheey°* and Hyunggon Park°*
Department of Electronic and Electrical Engineering, Ewha Womans University°,
Graduate Program in Smart Factory, Ewha Womans University*

요 약

본 논문은 복잡한 고차원의 네트워크 트래픽 데이터에 대해 특징 선택 기법 중 상호의존정보를 활용한 오토인코더 기반의 이상탐지 시스템을 제안하였다. 현대의 네트워크 트래픽 특성을 충분히 반영할 수 있는 데이터셋에 대해 상호의존정보 값이 클수록 중요한 특징으로 판단하고, 이에 따라 특징을 하나씩 추가하며 특징 수에 따른 이상탐지 오토인코더의 성능 변화를 확인하고 분석하였다. 특징 중요도 순서에 따라 특징을 추가하는 경우가 무작위 순서로 추가하는 경우보다 더 적은 특징만으로도 이상탐지 오토인코더의 성능을 향상시킬 수 있음을 확인하였다. 나아가 정답 레이블과의 상관관계에 그치지 않고, 선택하고자 하는 특징들 간 상관관계를 고려한다면 최적의 특징 조합을 구성할 수 있을 것으로 기대된다.

Ⅰ. 서 론

전통적인 컴퓨터 네트워크와 셀룰러 네트워크에서 소프트웨어 정의 네트워크(Software Defined Network, SDN) 및 사물인터넷(Internet of Things, IoT)에 이르기까지 네트워크 장치 수가 증가하고 그 구조가 복잡해짐에 따라 신속하고 정확한 네트워크 트래픽 이상탐지(anomaly detection)가 네트워크 시스템 유지 및 관리에 핵심적인 요소가 되었다. 현대의 네트워크 트래픽 이상 데이터는 정상 데이터와는 구분이 어려울 정도로 지능적으로 발전하고 있으며, 많은 수의 특징을 포함하는 고차원 데이터이다. 따라서 이상탐지 시스템 설계에 있어 특징 선택(feature selection) 기법을 통해 고차원 데이터를 효과적으로 다루는 과정이 요구되며, 복잡한 실제 네트워크 트래픽 데이터를 활용한 검증 과정이 필수적이다. 특징 선택은 중요하지 않거나 중복성이 높은 특징(feature)을 제거함으로써 전체 데이터를 대표할 수 있는 최적의 특징 조합으로 데이터를 정제하는 과정이다[1]. 대표적으로 확률 변수 간 비선형적인 관계를 반영할 수 있다는 특성으로 인해 효과적인 방법이라 알려진, 특징 간 정보량을 수치화한 상호의존정보를 활용하여 특징을 선택하는 상호의존정보(mutual information) 기반 특징 선택 기법이 있다[2].

네트워크 트래픽 이상탐지를 위한 데이터 전처리 연구와 더불어, 비지도 학습 기반의 심층 신경망인 오토인코더를 활용한 연구가 활발히 이루어지고 있다. 오토인코더는 데이터의 구조와 패턴을 파악하여 학습하는 비지도 학습 방식이기 때문에 보편적으로 알려지지 않아 예측하기 어려운 이상치에 대해서도 탐지 가능하다[1]. 특징 선택 기법을 통해 정제된 데이터를 오토인코더에 활용한다면, 대량의 네트워크 트래픽 데이터의 특성을 효율적으로 학습하여 탐지 정확도가 향상될 수 있으며, 오토인코더의 복잡도가 줄고 연산 처리 속도가 빨라져 신속한 이상탐지가 가능하다.

본 논문에서는 더욱 복잡해지고 정밀해진 고차원 네트워크 트래픽 데이터에 대해 상호의존정보를 이용해 특징 선택 과정을 거친 오토인코더 기반 이상탐지 시스템을 제안한다. 상호의존정보 값에 따라 특징을 추가하며 특징 수와 오토인코더의 이상탐지 성능 사이의 관계를 파악하고, 최근 발생하는 고차원 네트워크 트래픽 데이터의 속성을 반영한 데이터를 통해 상호의존정보를 이용한 오토인코더 기반의 이상탐지 시스템의 효율성을 검증한다.

Ⅱ. 상호의존정보를 이용한 오토인코더 기반 이상탐지 시스템

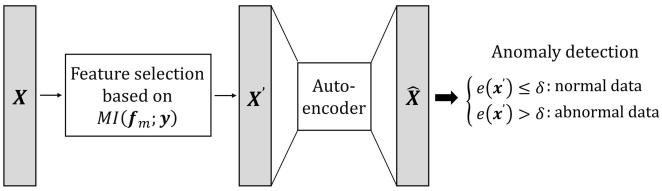

그림 1. 상호의존정보 기반 오토인코더를 이용한 이상탐지 순서도

본 논문에서는 그림 1과 같이 상호의존정보 값에 따라 특징을 추가하며 데이터의 이상여부를 판단하는 오토인코더 기반의 이상탐지 시스템을 제안한다. 네트워크 트래픽 데이터셋 $\boldsymbol{X} \in \mathbb{R}^{(M+1)\times N}$은 M개의 특징 $\boldsymbol{f}_m \in \mathbb{R}^N$ $(m = 1,2,\cdots,M)$ 을 갖는 N개의 데이터 $\boldsymbol{x}_n \in \mathbb{R}^M$ $(n = 1,2,\cdots,N)$와 레이블(label) $\boldsymbol{y} \in \mathbb{R}^N$로 구성되어 있다. m번째 특징 $\boldsymbol{f}_m$과 $\boldsymbol{y}$의 상호의존정보 $MI(\boldsymbol{f}_m;\boldsymbol{y})$는 다음과 같다.

$$MI(\boldsymbol{f}_m;\boldsymbol{y}) = \sum_{f\in \boldsymbol{f}_m}\sum_{y\in \boldsymbol{y}} p(f,y)\log\left(\frac{p(f,y)}{p(f)\,p(y)}\right) \qquad (1)$$

상호의존정보는 두 확률 변수 사이의 상호 의존성을 측정하는 지표로, 어떤 확률 변수가 다른 확률 변수에 대해 제공하는 정보량을 정량화한 값이며, $MI(\boldsymbol{f}_m;\boldsymbol{y})$는 데이터를 구성하는 $\boldsymbol{f}_m$과 $\boldsymbol{y}$의 상관관계(correlation)로 해석할 수 있다. MI가 큰 특징일수록 정상 데이터와 이상 데이터 구분에 관련성이 높다. 반면 MI가 0이면 $\boldsymbol{f}_m$과 $\boldsymbol{y}$는 독립적인 관계이다[3]. 따라서 $MI(\boldsymbol{f}_m;\boldsymbol{y})$를 $\boldsymbol{f}_m$의 특징 중요도 F_m이라 정의하고, F_m이 클수록 $\boldsymbol{f}_m$이 이상탐지에 미치는 영향이 크다고 판단한다.

특징 중요도 F가 가장 큰 k $(1 \le k \le M)$개의 특징으로 구성된 데이터셋 $\boldsymbol{X}' \in \mathbb{R}^{k\times N}$을 오토인코더의 입력으로 사용하며, 정상 데이터와 이상 데이터를 구분하기 위해 L_1-norm 기반 이상치 점수를 사용한다. n번째 데이터 $\boldsymbol{x}_n' \in \mathbb{R}^{k\times 1}$에 대한 이상치 점수 $e(\boldsymbol{x}_n')$는 다음과 같다.

$$e(\boldsymbol{x}_n') = \frac{1}{k}\sum_{i=1}^{k} |x'_i - \hat{x_i}| \qquad (2)$$

오토인코더의 출력 데이터 $\hat{\boldsymbol{x}}$과 특징 선택 과정을 거친 입력 데이터 $\boldsymbol{x}'$의 차이인 이상치 점수가 임계값 δ보다 크면 이상 데이터로 판단한다.

Ⅲ. 상호의존정보 기반 특징 개수에 따른 이상탐지 실험

제안한 시스템의 효율성을 검증하기 위해 무작위 순서로 특징을 추가하는 경우와 특징 중요도 F가 큰 특징부터 순차적으로 특징을 하나씩 추가하는 경우에 대한 실험을 진행한다. M개의 특징이 모두 포함될 때까지 선별 특징 개수 k를 증가시키며 오토인코더를 이용한 이상탐지를 반복한다. 각각 특징 중요도 순서와 무작위 순서의 특징 추가에 따른 오토인코더의 탐지 정확도를 확인한다.

첫 번째로 정상 데이터와 이상 데이터의 구분이 어려울 정도로 지능적으로 발전한 실제 네트워크 트래픽의 속성을 반영하기 위하여 인위적으로 생성한 고차원의 데이터셋을 사용하여 실험을 진행한다. 생성된 데이터셋은 가우시안 분포를 따르며 각 특징별로 정답 레이블과의 상호의존정보가 상이한 200차원 데이터이다. 정상 데이터와 이상 데이터의 평균은 각각 150 이상 200 이하, 170 이상 180 이하의 범위 내에서, 표준편차는 10 이상 30 이하의 범위 내에서 각 특징별로 무작위로 설정한다.

그림 2는 무작위 순서로 특징을 추가할 때와 특징 중요도 순서로 특징을 추가할 때의 이상탐지 정확도를 비교한 것이다. 무작위로 특징을 추가하는 경우, 특징 수가 증가함에 따라 성능이 점차 향상되어 모든 특징을 사용하였을 때 약 85%의 이상탐지 정확도에 도달하는 것을 관찰할 수 있다. 반면, 특징 중요도가 높은 순서로 특징을 추가하는 경우에는 특징 수가 매우 적을 때도 비교적 높은 정확도를 보이며, 정확도가 빠르게 증가하는 것을 볼 수 있다. 높은 중요도의 추가에 따라 약 98%의 정확도까지 빠르게 도달하며, 이후 중요도가 더 낮은 특징이 추가됨에 따라 정확도가 하락하여 약 85%로 수렴하는 것을 볼 수 있다. 실험 결과를 통해 상호의존정보 값을 기반으로 특징을 하나씩 추가하는 방식으로 특징 선택을 하는 경우 특징 중요도가 높은 일부 특징만 사용하는 것이 전체 특징을 모두 사용하는 것보다 이상탐지에 유리하다는 것을 알 수 있다.

두 번째로 실제 네트워크 환경과 유사한 환경을 구축하여 생성한 CSE-CIC-IDS 2018을 사용하여 실험을 진행한다. CSE-CIC-IDS 2018은 다양한 공격 유형의 네트워크 이상 데이터를 포함하는 데이터셋이다. 실험에는 전체 특징 중 IP 주소, port number와 같은 소켓 정보 및 protocol 그리고 중복되는 특징을 제외한 특징 76개만을 활용한다.

그림 3은 CSE-CIC-IDS 2018에 대한 특징 수에 따른 이상탐지 정확도이다. 특징 중요도 순서로 특징이 추가되는 경우가 무작위로 추가되는 경우보다 대체로 더 우수한 성능을 보인다. 2차 다항식 회귀(regression)를

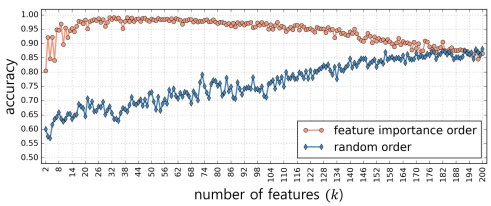

그림 2. 특징 추가 방법과 특징 수에 따른 인위적으로 생성한 데이터셋의 이상탐지 정확도

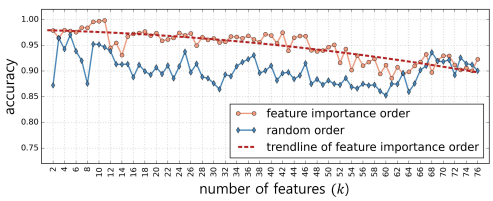

그림 3. 특징 추가 방법과 특징 수에 따른 CSE-CIC-IDS 2018의 이상탐지 정확도

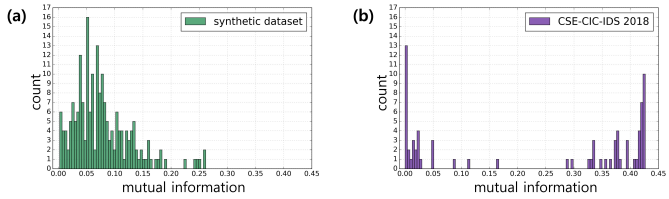

그림 4. 데이터셋별 상호의존정보 값 분포

통해 특징 중요도 순서에 따른 정확도 값의 추세선을 표시하였다. 특징 중요도 순서에 따라 특징을 추가하는 경우, 상호의존정보가 낮은 특징이 추가됨에 따라 이상탐지 정확도가 하락하는 경향성이 있다. 반면, 무작위로 특징을 추가하는 경우에는 특정한 경향성을 보이지 않는다.

첫 번째 실험과 두 번째 실험 결과의 차이는 데이터셋별 상호의존정보 값 분포에서 원인을 찾을 수 있다. 그림 4의 (a)는 인위적으로 생성된 데이터셋, (b)는 CSE-CIC-IDS 2018의 상호의존정보 값 분포를 나타낸 히스토그램이다. (a)는 상호의존정보 값이 상대적으로 작은 0부터 0.26까지 범위에 존재하는 반면, (b)는 32개의 특징은 0.16 이하의 상호의존정보 값을, 나머지 44개의 특징은 0.26 이상의 상호의존정보 값을 가진다. 따라서 CSE-CIC-IDS 2018을 사용한 실험에서는 무작위 순서로 특징을 추가할 때 인위적으로 생성된 데이터셋을 사용한 실험보다 상호의존정보 값이 큰 특징이 선택될 확률이 높으며, 특징별 상호의존정보 값의 편차가 크므로 뚜렷한 경향성이 나타나지 않는다. 특징 중요도 순으로 특징을 추가할 경우, 첫 번째 실험에서보다 두 번째 실험에서 처음으로 추가되는 특징의 상호의존정보 값이 크기 때문에 정확도가 증가하는 구간이 생략되고 대체로 감소하기만 하는 차이를 보인다. 또한 그림 3의 특징 중요도 순 실험에서는 정확도 곡선이 약하게 진동하는데, 이는 특징을 추가할 때 특징 간의 상관관계(correlation)를 고려하지 않았기 때문이다. 인위적 생성 데이터셋은 서로 다른 분포에서 샘플링한 독립적인 특징들로 구성되어 있지만, 실제 네트워크 트래픽 데이터는 '패킷이 전송된 시간'과 '전송된 패킷의 길이'와 같이 특성이 명확하고, 복잡한 특징 간 상관관계가 존재한다.

Ⅳ. 결 론

본 논문에서는 특징 중요도를 활용한 특징 선택을 거친 오토인코더 기반 이상탐지 시스템을 제안하고, 중요도에 따라 점진적으로 특징을 추가하며 탐지 성능 변화를 확인하였다. 특징 중요도 기반으로 특징을 선택하면 소수의 특징만으로도 뛰어난 탐지 성능을 낼 수 있다는 실험 결과를 활용하여, 최적의 특징 개수를 자동으로 선정하는 방법에 대한 후속 연구를 진행할 예정이다. 또한 특징 선택 시 특징과 레이블 간의 상관관계뿐만 아니라 특징들이 함께 공유하고 있는 정보량 등의 상관관계도 고려한다면 네트워크 트래픽 이상탐지 성능을 더욱 향상시킬 수 있을 것으로 기대된다.

ACKNOWLEDGMENT

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No. 2021-0-00739, 분산/협력AI 기반 5G+ 네트워크 데이터 분석 기능 및 제어 기술 개발)과 2020년도 한국연구재단의 지원(No. NRF-2020R1A2B5B01002528)을 받아 수행된 연구임.

참 고 문 헌

[1] S. Wang, et. al., "Machine Learning in Network Anomaly Detection: A Survey," *IEEE Access*, Vol. 9, pp. 152379-152396, 2021.
[2] J. R. Vergara, et. al., "A review of feature selection methods based on mutual information," *Neural Computing and Applications*, Vol. 24, pp. 175-186, 2014.
[3] A Realistic Cyber Defense Dataset (CSE-CIC-IDS2018) was accessed on Jan. 10, 2023. from https://registry.opendata.aws/c se-cic-ids2018.