응용 트래픽 분류를 위한 공개 데이터셋의 전처리 동향

장윤성, 최정우, 김용훈, 박재원, 김명섭*

고려대학교

{ brave1094, choigoya97, kyh9432242, 2018270614, tmskim* }@korea.ac.kr

Pre-processing Trends in Public Datasets for Application Traffic Classification.

Jang Yoon Seong, Choi Jeong Woo, Kim Yong Hun, Park Jae Won, Kim Myung Sup

Korea Univ.

요 약

네트워크 트래픽 분류는 네트워크 관리 분야의 핵심 기술로 이를 위해 다양한 공개 데이터셋이 활용되고 있다. 데이터셋의 특징과 연구 목적에 부합하도록 데이터셋을 전처리하는 것은 중요하며, 주요 전처리 과정에 대한 분석은 분류 모델의 성능에 도움을 줄 수 있다. 기존 연구에서는 대체로 비교 대상으로 선정한 연구의 전처리 과정을 고려하지 않고 모델의 성능만을 비교하는데, 객관적인 비교를 위해서는 동일한 전처리 과정을 통하여 마련된 데이터셋을 바탕으로 비교해야한다. 그러므로, 본 논문에서는 공개 데이터셋 UNSW-NB15, KDDCup99, NSL-KDD, CIC-IDS2017, ISCX VPN2016에 대하여, 적용된 전처리 과정을 비교하여 분석한다. 이를 통해, 기존 연구와의 객관적인 비교를 위한 전처리 과정의 가이드라인을 제시하고, 나아가 연구에 사용하는 데이터셋의 전처리 기법의 선택을 돕고자 한다.

Ⅰ. 서 론

트래픽 분류는 네트워크 보안, 성능 최적화, 서비스 품질 관리 등 다양한 네트워크 관리 및 보안 목적을 위해 중요한 역할을 하는 분야이다. 이러한 트래픽 분류는 최근 딥러닝 기술을 적용한 형태로, 복잡한 패턴을 스스로 학습하고 더 높은 정확도와 효율성을 제공할 수 있도록 연구되어 왔다. 이러한 상황에서 데이터셋은 딥러닝 기반 트래픽 분류 연구에 있어 핵심적인 구성 요소 중 하나로, 모델 학습 및 평가를 위한 핵심 자원이다. 연구자들은 데이터셋 수집의 어려움 때문에, 다양한 공개 데이터셋에 특정 전처리 과정을 적용하여 트래픽 분류 모델을 학습시키는 데에 사용한다. 각 연구에서는 데이터셋마다 공통적으로 수행되는 전처리 과정이 드러나기도 하는데, 본 논문에서는 최근 10년간의 연구에서 적용한 전처리 과정을 바탕으로 데이터셋 별 주요 전처리 기법을 비교 분석하여 제공한다. 이를 통하여, 응용 트래픽 분류를 위한 데이터셋의 전처리 과정의 가이드라인을 제시하고, 연구 데이터셋의 확보를 돕고자 한다.

Ⅱ. 본론

데이터셋은 트래픽 분류 모델을 학습시키는 데에 핵심적인 역할을 하는 딥러닝 요소이다. 전처리는 모델의 특징에 따라 데이터의 형식을 변환하고, 데이터 불균형도를 조정하는 등 다양한 이유로 사용된다. 이전 연구에서 저자(본 논문 작성자)는 데이터셋 별 전처리 기법에 대하여, 데이터셋 정제, 표준화, 특징 추출, 분할의 4단계로 구분하여, 데이터셋이 각 단계별로 어떠한 전처리 과정을 거쳐 연구 데이터셋으로 확보되는지를 비교하여 제시한다 [1]. 본 논문에서는 대체로 연구에서 고안하는 모델을 제시하는 특징 추출 단계를 제외한 나머지 단계의 데이터셋 별 주요 전처리 기법을 3단계로 구분하여 비교하고 분석하여 제시한다. 트래픽 분류를 위한 데이터셋 별 주요 전처리 기법은 다음과 같다.

본 과제(결과물)는 2023년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과이며 (2021RIS-004), 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(00235509, ICT융합 공공 서비스•인프라의 암호화 사이버위협에 대한 네트워크 행위기반 보안관제 기술 개발)임

1) UNSW-NB15 : 정제 단계에서는 One-hot Encoding, Label Encoding 기법이 가장 높은 빈도를 보인다. 이는 분류 모델에 적용하기 위하여 데이터 형식을 범주형 데이터에서 정수형 데이터로 변환하기 위해 사용되었다. 또한 데이터의 불균형도 IR(Imbalance Ratio)의 문제를 해결하기 위한 오버샘플링 기법 SMOTE가 적용되기도 한다. 불균형도를 의미하는 IR은 다수 클래스의 샘플수를 소수 클래스의 샘플수로 나눈 값이며, 0에 가까울수록 데이터의 샘플 수가 균형을 이루는 것이고, 값이 클수록 불균형도가 크다는 의미를 가진다. Guo 등은 최근 연구에서 대부분의 연구에서 고려되지 않은 점을 두 가지 언급한다 [2]. 첫 번째로는, 인터넷 트래픽이 자연스럽게 불균형한 분포를 이룬다는 것과, 두 번째로는 대부분의 연구에서 주로 클래스의 불균형을 고려하지 않는다는 것이다. 데이터의 불균형도가 클수록, 모델의 편향학습이나, 과적합의 문제가 발생할 수 있기 때문에, 이에 대한 전처리가 필요하다는 것이다. 그 외에 누락 값 처리를 위해 사용된 Linear Interpolation Method가 있었으나, 자주 적용되지는 않았다. 표준화 단계에서는 Min-Max Normalization, Z-score Standardization, L2 Normalization 의 순서로 높은 빈도를 보인다. 범주가 다른 특징들에 대하여 척도를 동일하게 함으로써, 데이터 간 비교와 분석이 더 용이하도록 하고자 적용되는 전처리 기법들이다. 데이터의 분할 단계에서는 Train set : Test set의 비율을 7:3, 8:2로 결정하는 경우가 가장 많았고, k-fold 방식을 사용하거나, 데이터셋이 수집된 16시간, 15시간으로 나눈 연구도 있었다. 이와 같은 방식의 분할방법도 고려해볼 수 있다.

2) KDDCup99 : 정제 단계에서는 One-hot Encoding이 가장 높은 빈도를 보이며 사용되었다. 시계열 데이터에서 크기가 다른 샘플을 동일한 크기로 맞추기 위해, 0으로 채우는 Zero Padding 기법이 사용되었지만 자주 적용되지는 않았다. 표준화 단계에서는 Min-Max Normalization, Z-score Standard -ization 기법이 적용되었다. 데이터 분할 단계에서는 6:4, 3:2, 17:3의 비율과 10-fold로 적용하기도 하였다.

데이터셋	정제		표준화	분할
UNSW-NB15	변환	**One-hot Encoding (11)** **Label Encoding (5)**	**Min-Max Normalization (11)** **Z-score Standardization (7)** L2 Normalization (1)	**7:3 (9)** **8:2 (3)** 5-fold (1) 10-fold (1) 3:1:1 (1) 66:31:3 (1) 16h:15h (1)
	IR처리	**SMOTE (2)**		
	그 외	Linear Interpolation Method (1)		
KDDCup99	변환	**One-hot Encoding (5)** Label Encoding (1)	**Min-Max Normalization (5)** **Z-score Standardization (4)**	**6:4 (2)** 10-fold (1) 3:2 (1) 17:3 (1)
	IR처리	-		
	그 외	Zero Padding (1)		
NSL-KDD	변환	**One-hot Encoding (19)** **Label Encoding (9)**	**Min-max normalization (20)** **Z-score standardization (7)** Log normalization (2) L2 normalization (1)	**17:3 (11)** **8:2 (5)** 7:3 (3) 4-fold (1) 7:1:2 (1) 66:31:3(1)
	IR처리	**SMOTE (2)** **ADASYN (1)**		
	그 외	-		
CIC-IDS2017	변환	**One-hot Encoding (5)** Label Encoding (1) Word2vec Method (1)	**Min-max normalization (8)** **Z-score standardization (4)** 0.1-0.9 normalization (1) Batch normalization (1)	**4:1 (5)** **7:3 (2)** 6:2:2 (1) 17:3 (1) 4-fold (1) 18:1:1 (1)
	IR처리	**SMOTE (2)** **Random Oversampling (2)**		
	그 외	-		
ISCX VPN2016	변환	**One-hot Encoding (2)**	**Min-Max normalization (4)** Batch normalization (1)	3:1:1 (1) 10-fold (1) 5-fold (1) 8:1:1 (1) 7:1:2 (1)
	IR처리	-		
	그 외	-		

표 1. 데이터셋 별 주요 전처리 기법

3) NSL-KDD : 정제 단계에서는 One-hot Encoding, Label Encoding 기법이 가장 높은 빈도를 보였다. NSL-KDD 데이터셋 또한 불균형이 심하기 때문에, 이를 처리하기위한 SMOTE, ADASYN과 같은 오버샘플링 기법이 적용되기도 하였다. SMOTE는 소수 클래스의 샘플을 생성하여 데이터셋을 균형감 있게 만드는 기법이고, ADASYN은 소수 클래스의 밀도에 따라 생성되는 샘플의 수를 조절하여 생성하는 기법이다. 표준화 단계에서는 Min-Max Normalization, Z-score Standardization, Log normalization, L2 Normalization 등의 방식이 사용되었다. 데이터의 분할 단계에서는 Train set : Test set의 비율을 17:3로 결정하는 경우가 가장 많았는데, k-fold 방식을 사용하거나, 데이터셋이 수집된 16시간, 15시간으로 나눈 연구도 있었다. 이와 같은 분할 비율도 고려해볼 수 있다.

4) CIC-IDS2017 : 정제 단계에서는 데이터 변환을 위해 One-hot Encoding 기법이 가장 많이 적용되었지만, Label Encoding과 Word2vec Method과 같은 변환 방식도 고려해 볼 수 있다. 또한 이 데이터셋의 매우 심각한 불균형도를 해결하기 위해, SMOTE, Random Oversampling과 같은 전처리 기법이 적용되었다. Random Oversampling 기법은 소수 클래스 데이터의 일부 데이터를 복사하여 기존 소수 클래스에 추가함으로서, 소수 클래스의 샘플 수를 증가시키는 오버샘플링 기법이다. 또한, 표준화 단계에서는 Min-Max Normalization, Z-score Standardization, 0.1-0.9 normalization, Batch Normalization 기법과 같은 다양한 표준화 전처리가 적용되었다. 데이터의 분할은 주로 4:1의 비율로 이루어졌고, 최근에는 Train : Validation : Test 의 비율로 분할하는 방식도 사용되었다.

5) ISCX VPN2016 : 정제 단계에서는 주로 연구자의 판단에 의해, 헤더를 제거하거나, 노이즈를 제거하는 등의 다양한 전처리 과정이 포함되었으나, 연구 별로 다양하기 때문에, 포함하지 않았다. 표준화 단계에서는 Min-Max Normalization이 가장 높은 빈도를 보였고, 분할 단계에서는 3:1:1, 8:1:1, 7:1:2와 같은 Train : Validation : Test 비율로 분할하는 경우가 가장 많았고, k-fold방식으로 분할하는 경우도 있었다. 이는 모델의 일반화 성능을 평가하기 위해 사용된 방식으로, 연구에서 불균형으로 인한 과적합 문제에 대한 대안으로 사용되는 기법들이다.

Ⅲ. 결론

본 논문에서 공개 데이터셋 UNSW-NB15, KDDCup99, NSL-KDD, CIC-IDS2017, ISCX VPN2016에 대해 최근 연구에서 적용하는 주요 전처리 기법을 비교하고 분석함으로써, 향후 연구에서 트래픽 분류를 위한 공개 데이터셋의 전처리를 지원하고자 한다. 분류 모델에 적용시키기 위한 단계로 One-hot Encoding과 Label Encoding의 데이터 변환, SMOTE와 ADASYN의 불균형도 문제 해결, Min-Max Normalization, Z-score Standardization의 표준화 단계를 거쳐 데이터셋을 확보하는 것이 데이터셋의 종류와 상관없이 높은 빈도를 보이며 수행되었다. 데이터셋의 분할은 Train : Test로 비율이 주로 사용되었고, Train : Validation : Test 분할 방식과, K-fold 분할 방식을 고려해볼 수 있다. 이러한 가이드라인을 통하여, 기존의 전처리 과정과 분류 모델을 공개하여 재현성이 확보된 연구에 한하는 동일 조건 하의 성능 비교가 가능해지고, 향후 연구에서 모델을 학습시킬 연구 데이터셋을 확보할 수 있다. 마지막으로 후속 연구과제로서 데이터 불균형도의 해결을 위한 연구와 일반화 성능을 고려한 데이터셋 분할 비율에 대한 연구를 제시한다.

참 고 문 헌

[1] 장윤성, 박지태, 백의준, 김주성, 이대국, 김명섭, "딥러닝 기반의 응용 프로그램 트래픽 분류를 위한 데이터셋 사용 및 전처리 방법", KNOM Review, 2023, Vol.26, No.2, pp.33-48, Dec 2023.

[2] Anik Vega Vitianingsih, Zahriah Othman, Safiza Suhana Kamal Baharin, Aji Suraji, Anastasia Lidya Maukar, "Application of the Synthetic Over-Sampling Method to Increase the Sensitivity of Algorithm Classification for Class Imbalance in Small Spatial dataset.", Internationlal Journal of Intelligent Engineering & Systems, 2022, Vol. 15 Issue 5, p676-690, August 2022.