JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 25, No. 11, pp. 3273-3280, Nov. 2024

Check for updates

레이블 교체 방식 기반 지식그래프 부정 샘플링 및 지식 그래프 학습 모델

이 중 훈[1] · 오 승 민[2] · 김 광 기[3] · 한 민 수[4] · 김 진 술[5*]
[1]전남대학교 지능전자컴퓨터공학과 석사과정
[2]전남대학교 지능전자컴퓨터공학과 박사과정
[3]나사렛대학교 스마트미디어트랙과 교수
[4]아스타나IT대학교 컴퓨터 및 데이터 과학과 교수
[5]전남대학교 지능전자컴퓨터공학과 교수

Negative Sampling based on Label Swap for a Knowledge Graph

Junghoon Lee[1] · Seungmin Oh[2] · Kwangki Kim[3] · Min-Soo Hahn[4] · Jinsul Kim[5*]

[1]Master's Department of Intelligent Electronics and Computer Engineering, Chonnam University, Gwang-ju 61186, Korea
[2]Ph.D Department of Intelligent Electronics and Computer Engineering, Chonnam University, Gwang-ju 61186, Korea
[3]Professor Department of IT Artificial Intelligence, Korea Nazarene University, Chungcheongnam-do 31172, Korea
[4]Professor, Department of Computational and Data Science, Astana IT University, Astana 010000, Kazakhstan
[5]Professor, Department of Intelligent Electronics and Computer Engineering, Chonnam University, Gwang-ju 61186, Korea

[요 약]

현대 네트워크 인프라가 복잡해짐에 따라 고도화된 기술과 대응책이 필요해지고 있다. 그러나 기존 장애 분석은 주로 특정 임곗값 반응 등을 통해 이상 상황을 예측하며, 이러한 접근 방식은 네트워크 구조적 정보를 충분히 활용하지 못한다. 이를 위해 시계열 데이터의 지식 그래프 변환을 적용하여 네트워크 관계 정보를 학습하는 방법을 제안한다. 모델의 일반화 성능을 향상하기 위해 부정 샘플링을 적용하였으며, 하여 각 노드의 특성을 업데이트하고, 관계를 학습하여 성능을 개선하였다. 이를 통해 데이터 빈도수가 높게 나타나는 '정상' 상태의 편향증상을 완화하였고, 모델을 더 다양한 시나리오에 노출하여 성능을 향상했다. 이 방법을 통해, 유사 네트워크 장애 구별 능력을 크게 향상하였으며, 즉각적 원인 외의 고려 요인들을 다루어 개선하였다.

[Abstract]

As modern network infrastructures become increasingly complex, the need for advanced technologies and countermeasures is growing. Traditional fault analysis methods often rely on specific thresholds to predict anomalies, which do not fully leverage the structural information within the network. To address this limitation, we propose a method that converts time-series data into a knowledge graph to learn network relationship information. Negative sampling is employed to enhance the model's generalization performance by updating the features of each node and learning their interrelationships. This approach reduces the bias toward the 'normal' state, which is more prevalent in the data, and improves performance by exposing the model to a wider range of scenarios. Consequently, the ability to distinguish similar network faults has significantly improved, allowing for the consideration of various factors beyond immediate causes, leading to overall operational enhancements.

색인어 : 지식 그래프, 부정 샘플링, 네트워크, GCN, 데이터 증강
Keyword : Knowledge Graph, Negative Sampling, Network, GCN, Data Augmentation

http://dx.doi.org/10.9728/dcs.2024.25.11.3273

Received 30 August 2024; **Revised** 02 October 2024
Accepted 18 October 2024

***Corresponding Author; Jinsul Kim**
Tel: +82-62-530-1808
E-mail: jsworld@jnu.ac.kr

Ⅰ. 서 론

현대의 네트워크 관리에서는 다양한 종류의 장애를 신속하고 정확하게 감지하는 것이 중요하다. 그러나 상황에 따른 네트워크 구조의 특수성과 공간적 정보를 이해하여 예측하기란 쉽지 않다. 그림 1에 의하면, 실제 네트워크에 대한 A와 B의 사례에서만 보더라도 특정 디바이스 또는 구성 인터페이스가 상/하위에 배치됨에 따라 오류 발생 위치의 중요도 및 오류의 종류에 큰 차이점이 생김을 알 수 있다. 이를 통해서 네트워크의 이상 탐지 시 네트워크의 구조적 정보를 이해한 것이 중요함을 알 수 있지만, 대부분의 네트워크 이상 탐지를 위한 학습 데이터에는 이러한 공간적 정보를 제공하는 경우는 드물며, 보통 각 디바이스 또는 인터페이스에서 제공하는 통신상태 및 통신 완료 속도, 패킷 등의 시스템의 활동성 및 계측적인 데이터에 치중하고 있는 경우가 많다. 또한, 이러한 시계열 기반의 데이터 적 특징은 데이터 대부분이 특정한 상태(일반적인 예시로는, normal 한 상태)일 때, 해당 레이블의 빈도가 높기 때문에 데이터 학습 시 모델의 알고리즘에 편향 문제를 일으킬 가능성이 있다[1]. 이러한 문제들을 해결하기 위한 방법으로서, 최근 연구에는 네트워크 시계열 데이터를 기반으로 하는 지식 그래프화 연구가 진행되고 있다. 일례로는, 지식 그래프의 엔터티와 관계 간의 복잡한 고차원 관계를 저차원 연속 벡터 공간에 내장하여 고차원 복잡한 관계를 더 쉽게 찾고 계산할 수 있도록[2] 하는 방법이 있다. 본 논문에서는 이러한 방법을 통해 기존의 문제점들을 극복하고 현 방식의 학습 효율을 개선하기 위해 네트워크 시계열 데이터를 지식 그래프로 변환하고, 이를 학습하는 방법으로서 그래프 컨볼루션 네트워크(GCN; Graph Convolution Network)모델을 적용하였다. 또한, 기존의 편향된 네트워크 데이터 특성을 개선하고 학습의 효율을 높이기 위한 방법으로서 부정 샘플링(Negative Sampling)[3]을 도입하였으며, 나아가 모델 성능의 개선을 위해 기존 부정 샘플링 방법에서 새로운 접근 방식인 레이블 교체(Label Replacement) 기법을 제안하여 그 효과를 비교하였다. 이러한 접근을 통해 네트워크 장애 예측의 정확도를 높이고, 더 나은 모델 일반화를 실현하고자 하였다.

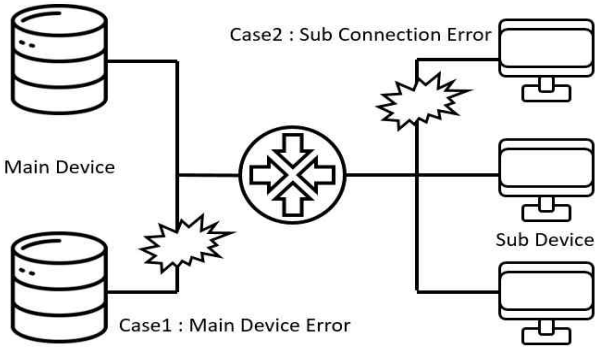

그림 1. 네트워크 장애 시나리오 케이스
Fig. 1. Case of network failure scenario

최종적으로, 본 논문에서 연구를 통해 해결하고자 하는 내용은 다음과 같다.

1) GCN을 통해 네트워크 구조의 복잡한 구조적 정보를 바탕으로 네트워크 장애를 효과적으로 학습할 수 있는가?

- GCN은 그래프 형태의 데이터에서 복잡한 구조적 관계를 학습할 수 있는 능력을 지녔으며, 이를 통해 네트워크의 토폴로지와 장애 발생 패턴을 이해하고, 네트워크 내에서 발생할 수 있는 다양한 유형의 장애를 예측하도록 연구를 수행하였다.

2) 협업 필터링 기법으로 유사 네트워크 장애를 예측할 수 있는가?

- 협업 필터링은 지식 그래프 기반의 기존 추천시스템에서 널리 사용되는 기법으로, 유사한 사용자 행동을 기반으로 추천을 생성하는 원리이다. 이점에 착안하여, 네트워크 장애 예측에 적용하여 이에 대한 유사한 장애 형태를 보이는 네트워크 상황을 예측할 수 있는지를 탐구하였다.

3) 제안된 모델이 기존의 네트워크 장애 분석 방식에 비해 성능 향상을 제공하는가?

- 기존의 시계열 탐지 기반 학습 데이터 세트 방식과 비교하여 더 높은 정확도와 효율성을 보이는지를 평가하였다.

따라서, 본 연구는 네트워크 형태의 시계열 데이터를 기반으로 한 지식 그래프의 구축을 진행하고, 해당 지식 그래프의 학습을 고도화하여 이상 분류 모델의 성능을 향상하고자 한다.

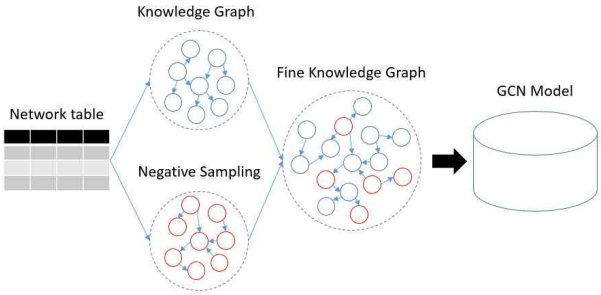

그림 2. 네트워크 시계열 데이터의 지식 그래프 기반 레이블 교체 방식 부정 샘플링 기법
Fig. 2. Label swap based negative sampling for network temporal data

Ⅱ. 관련 연구

기존의 네트워크에 대한 구조적 정보의 활용과 데이터의 편향성을 개선하기 위한 다양한 연구들이 진행되어 왔다. 예를 들어, 네트워크의 구조적 정보를 모델이 학습하기 위한 방법

으로서 기존의 시계열 데이터를 기반으로 하는 네트워크에 대한 연구로는 “Training convolutional networks with noisy labels”[4] 및 “Deriving validity time in knowledge graph.”[5]가 있고, 데이터의 성능개선을 증가하는 방법을 제시한 부정 샘플링 기반 연구로는 SMOTE(SMOTE, Synthetic minority over-sampling technique)[6]와 같은 사례가 있다. 그러나 해당 사례는 편향된 데이터를 기반으로 부정 샘플링을 생성할 경우, 부정 샘플링 또한 편향되게 생성된다는 문제가 있으며, 이러한 문제를 해결하는 방법으로 네트워크의 부정 샘플링 내 레이블을 교체하는 방식을 활용한 레이블 교체 기법을 제안하고자 한다. 해당 방법은 지식 그래프에서 부정 샘플링을 생성할 때 기존의 랜덤 생성 방식을 기반으로 하지만, 원본 데이터와 유사한 구조를 유지하기 위해 지식 그래프의 트리플 구조를 활용하여 전체 노드중 출발 노드와 도착 노드에 해당하는 노드끼리 분류를 진행하였고, 각각의 노드 군에서 1개씩 랜덤하게 꺼내 매칭 하는 것으로 원본 데이터의 형태와 유사하지만 오답인 부정 샘플링 데이터 세트를 생성하였다. 이후 해당 데이터세트에서 데이터 적 편향성을 위해, 해당 샘플 데이터를 기반으로 레이블에 해당하는 값을 랜덤하게 바꾸는 방식을 적용하여 실험을 진행하였다.

2-1 지식 그래프(Knowledge Graph)

지식 그래프란 그래프로 지식을 표현한 데이터 형태 중 하나로, 대상에 대한 관계성을 표현하기 위해 객체(Entity)를 정점(Vertex), 관계(Relation)를 간선(Edge)으로 표현한 것을 말한다. 일반적으로 지식 그래프는 테이블형 데이터 구조와 비교되는데, 관계형 데이터베이스와 지식 그래프의 가장 큰 차이점은 바로 확장성에 있다. 기존의 관계형 데이터베이스는 서로 다른 구조를 가진 테이블 간의 관계를 단순히 참조 형식으로 연결해야 하지만, 지식 그래프는 두 대상 간의 관계를 직접적으로 표현하여 데이터를 쉽게 연결할 수 있다. 이러한 특성 덕분에, 지식 그래프는 다양한 데이터 간의 관계를 보다 직관적이고 유연하게 탐색할 수 있다.

1) GCN 소개

GCN[13]란 각 노드의 특성을 그 주변 노드의 특성과 결합하면서 노드의 새로운 특성을 생성하는 것으로, 일반적으로 이미지 처리 과정에서 사용되는 컨볼루션 과정을 그래프 데이터에 적용한 사례이다. 그림 2에 의하면, 지식 그래프에는 두 가지의 데이터를 추출할 수 있으며, 첫 번째는 노드에 대한 Feature 정보를 가지고 있는 특징 행렬과 노드의 순서별로 다른 노드와의 연결 정보를 담고 있는 인접행렬이다.

지식 그래프 그래프 형태의 데이터에 컨볼루션과정을 적용할 경우, 그래프가 여러 노드를 이어주는 간선으로 연결되어 있으므로 하나의 노드에서도 지역적인 특성(Local Feature)을 추출할 수 있으며, 또한 필터가 전체 데이터에 같이 적용되기 때문에, 각 노드의 위치에 상관없이 처리되는 공간적 불변성의 특징을 가지게 되어 연관성 높은 데이터 간의 광범위한 패턴 정보를 학습할 수 있다.

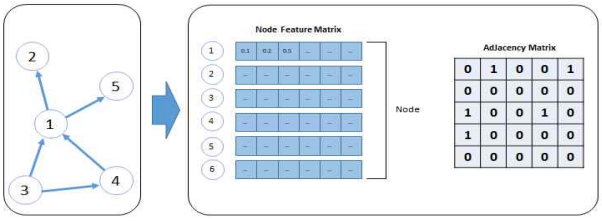

그림 3. GCN모델의 입력 데이터 생성 방식
Fig. 3. Way of creating GCN model’s input data

2) GCN의 작동 방식

- Aggregation(이웃 집계)[7] : GCN의 첫 번째 단계로, 이미지 데이터에서 feature를 집계하는 과정처럼 지식 그래프 내에 존재하는 한 노드의 feature와 그 노드와 직접 연결된 “이웃”의 노드 특성을 함께 집계하는 단계이다. 그러나 그래프의 Aggregation단계는 이미지 데이터와 다르게 구조적으로 다양한 관계 데이터를 포함하고 있어서 이러한 정보에 크게 영향을 받는다는 차이점이 존재한다. 그림3에서 그래프로부터 생성 가능한 두 가지 정보인 인접행렬(Adjacency Matrix)과 노드 특징 행렬(Node Feature Matrix)을 확인할 수 있다. 인접행렬 정보는 하나의 노드를 대상으로 인접한 노드의 연결 방향을 기록한 행렬이다. 노드 특징 행렬은 노드가 가진 feature를 담고 있는 행렬정보이다. 이 두 정보를 기반으로, 모델은 인접행렬 정보를 통해 조회 순서에 있는 노드의 연결 인접 정보를 확인하고, 노드 특징 행렬을 통해 노드의 특징이 연결된 노드에 어떤 영향을 부여하는지를 확인하게 된다. 이 과정에서 가중치 행렬을 사용하여 이웃의 정보와 대상 노드의 정보를 합산하여 최종적으로 새로운 노드의 특성 표현으로 생성된다. 이를 통해 각 노드는 자신의 이웃들로부터 수집한 정보와 자신의 정보를 결합하여 더 의미 있는 표현을 생성한다.
- Non-linear Transformation(비선형 변환)[8] : GCN 모델에 비선형성을 추가하여 표현력을 높이는 단계로서, 활성화 함수 ReLU를 사용하여 그래프 데이터의 복잡성과 다양한 관계를 효과적으로 학습한다.
- Hierarchical Struction(계층적 구조)[9] : GCN 모델에 계층적 구조를 적용하여, 레이어를 거칠 때마다 이웃 노드의 범위를 넓히고 지역적 특징을 전역 정보로 변환하는 역할을 수행한다. 결과적으로 이를 통해 노드의 특성을 주변 노드와 결합하고, 여러 레이어를 통해 전체 그래프의 전역 정보를 학습하게 된다.

3) 부정 샘플링

부정 샘플링은 “긍정”적인 데이터 이외에도, 부정의 사례데이터를 생성하여 학습시키는 일종의 증강데이터를 의미한다.

특히 지식 그래프 학습 시에 부정 샘플링은 중요한 역할을 하는데, 부정 샘플링을 학습함으로써 모델이 긍정적인 경우와 부정적일 때 모두를 구분하여 모델의 정교함을 향상할 수 있기 때문이다. 단, 부정 샘플링의 품질이 낮으면 잘못된 사례를 학습하게 되어 올바르게 이해하지 못하게 될 수 있다. 따라서 정교한 부정 샘플링의 생성은 매우 중요하며, 이를 위해 부정 샘플링의 생성에 대한 다양한 기법들이 제시되고 있다[10].

다음 그림 4.를 예시로 들면, 지식그래프 노드 [1, 2, 3, 4, 5]에 대해 참에 대한 노드간의 연결 정보가 Positive Relation의 영역으로 추출해냈음을 확인할 수 있다. 하지만 추가적으로, 해당 지식그래프에 참이 아닌 노드간의 연결 정보로 [2, 3], [2, 5], [5, 3]...와 같이, 기존의 참의 샘플링 정보에 대해 head, 또는 tail노드의 정보를 인위적으로 바꿔 참이 아닌 값으로 바꾸는 방식이다.

이 방식을 통해 모델은, 참의 샘플링 정보를 통해 손실 함수가 1에 가까운 확률이 되도록 학습하지만 반대로 부정 샘플링을 통해 학습한 정보는 0이 되도록 학습 하게 된다.

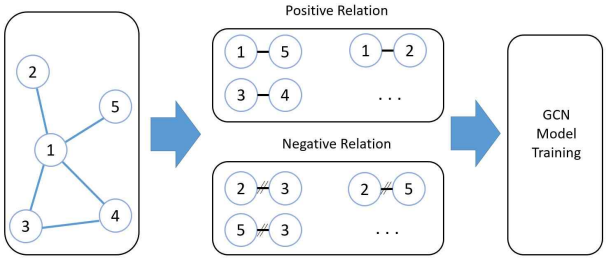

그림 4. 부정 샘플링 생성 방식
Fig. 4. How to create negative sampling

4) 레이블 교체 부정 샘플링 기법

본 논문에서 제시하고자 하는 부정 샘플링의 방법은 지식 그래프 기반의 네트워크의 이상 탐지 분야에 최적화하려는 방법으로, 기존의 방식들인 무작위 생성 또는 연관성 기반의 방식이 아닌 노드에 저장되는 레이블 정보를 교체하는 방법을 활용하였으며, 기존의 방식과는 다음의 차이가 있다.

- 레이블 정보를 활용 : 기존의 방식들은 노드 자체의 정보를 교체하여 부정 샘플링의 생성을 수행하였지만, 본 방법은 노드의 대부분 정보는 유지하되, 학습에 사용되는 레이블 정보를 랜덤하게 교체하였다.
- 온톨로지 구조의 유지 : 부정 샘플링을 생성 시 노드의 시작, 도착에 대한 범위를 고려하여 원본 데이터와 유사하면서도 새로운 데이터를 생성하였다.

Ⅲ. 실험 설계

3-1 참고문헌

GCN모델을 설계하여 3가지 종류의 데이터를 학습하여 성능을 비교하였다. 첫 번째 모델은 원본 데이터를 모델에 투입하였고, 두 번째 모델은 출발, 도착 노드에 대한 연결성을 고려한 원본과 유사한 부정 샘플링 데이터를 투입하였다. 마지막 모델은 본 논문에서 제안하는 레이블 변경을 활용한 부정 샘플링을 투입하여 학습에 사용하였다.

3-2 데이터 설명

본 실험을 위해 사용된 데이터는 BNN-UPC에서 Graph Neural Networking Challenge 2023에서 제공한 gnnet-challenge2023 데이터셋[11]의 네트워크의 디바이스 및 인터페이스의 구조에 대한 이상 분류용 학습 데이터를 사용하였으며, 949개에 대한 지표를 총 460번 측정한 데이터이다. 각 지표는 시간, y_true(fc)(정답 레이블), 그리고 디바이스 또는 인터페이스의 네트워크 관련 계측값에 대한 정보를 담고 있으며 약 4.6MB의 데이터를 담고 있다. (예시 : statistics.in-octets_value, statistics.out-unicast-pkts_value 등) 해당 데이터의 전처리를 위해 결측치의 제거 및 정규화 과정을 수행하였고, 총 측정 횟수를 늘리기 위해 데이터 스케일링 증강을 적용하여 [0.9, 0.8, 1.1, 1.2]에 대한 스케일을 적용하여 5배로 데이터의 수를 늘렸다.

3-3 부정 샘플링 생성방법

데이터의 편향성 문제를 해결하기 위해, 본 실험에서는 지식 그래프를 구성하는 트리플의 구조인 출발과 도착 노드의 개념으로 전체 노드를 분류를 진행하였다. 실험을 진행하기 위해 총 두 번의 부정 샘플링을 진행하였는데, 일반적인 부정 샘플링에 해당되는 비교군은 각각의 노드군에서 랜덤한 샘플링을 취해서 최소한의 유사성을 적용한 부정 샘플링 데이터를 생성하였다. 두 번째로 생성된 부정 샘플링은 본 논문에서 제안하는 방법인 레이블 교체 기반 부정 샘플링을 적용하였으며, 이는 랜덤 생성된 부정 샘플링에서 레이블에 해당하는 요소를 랜덤한 값으로 교체하여 다른 정답으로 교체하는 과정을 적용하였다. 이 데이터를 생성하는 과정에서, 랜덤으로 생성시 원래 지식 그래프의 정답 트리플과 중복되는지 확인하는 과정을 거쳐 총 200개의 일반 부정 샘플링과 레이블 교체기반 부정 샘플링을 생성하여 실험에 사용하였다.

3-4 그래프 컨볼루션 네트워크(GCN) 모델

증강한 데이터 및 부정 샘플링은 지식 그래프를 기반으로 생성되었으며, 따라서 노드 기반의 정보를 받아들이고 학습할 수 있는 모델로 GCN을 채택하였다. GCN의 핵심 목표는 그래프의 노드간 연결구조와 노드 자체의 특성을 고려하여 노드가 표현하는 연결 구조를 학습하는 것이다. 이를 위해 그래프 내의 노드간의 연굴 구조와, 노드 자체에 대한 특성을 고려하여 노드의 표현을 학습한다. 최종적으로, 전처리 된 데이

터 및 증강데이터를 컨볼루션 기반의 GCN 모델에 통과시키는 과정을 거치게 되며 이를 위해 아래의 수식을 통해 그래프의 노드를 통해 다음 번째의 노드를 예측하게 된다.

$$H^{(l+1)}=\sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}D^{-\frac{1}{2}}H^{l}W^{l}) \quad (1)$$

여기서 H는 레이어의 노드 표현(즉 l 및 l+1은 현재의 레이어노드 및 그다음의 레이어 노드)이며, D는 A의 대각행렬, W는 현재 번째의 레이어에 대한 학습 가능한 가중치 행렬을 의미하고, σ는 모델의 비선형성 추가를 위한 활성화 함수(ReLU)를 뜻한다.

3-5 손실 함수(Loss function)

손실 함수는 실제 레이블과 모델이 예측한 확률 분포 간의 차이를 측정하는 데 사용하며, 이를 통해서 각 노드 레이블의 예측에 대한 오차를 계산하는 데 사용한다. 여기서는 Cross Entropy Loss function을 사용하여 모델의 성능을 평가하였다. 해당 함수는 모델이 예측한 확률 분포와 실제 레이블의 분포 차이를 계산하여, 그 손실 최소화를 하도록 모델을 개선한다. 이에 대한 수식은 식 (2)에 표현되어 있다.

$$L=-\sum_{c=1}^{M}y_{o,c}\log(p_{o,c}) \quad (2)$$

여기서 M은 예측할 클래스의 수, y는 정답을 분류하기 위한 지시자(1 or 0), p는 모델이 예측된 확률, o와 c는 관측한 값 및 클래스 c를 의미한다. 따라서, 모델이 예측한 관측 대상이 c일 확률에 대해 로그를 취하여 정답에 가까울 때 0에 가까워지고 오답일 경우 음의 무한대로 커지도록 한다. 이후 해당 분류가 정답이면 1을 곱하고 아니면 0을 취하여 버린 다음, 해당 값들을 모두 취합하여 음수를 적용함으로써 최종적으로 모델의 성능을 평가하는 데 사용한다.

Ⅳ. 결과 해석

4-1 Confusion Maxtrix 분석

첫 번째 평가지표는 GCN모델에 부정 샘플링 없이 전처리한 데이터만을 학습한 결과의 Confusion Matrix이다. 이 데이터를 확인하였을 때, Confusion matrix 상에서 네트워크의 이상분류 중 normal에 해당하는 0은 가장 높은 정답률을 보이고 있으며, 이는 데이터상에서 이상분류 상황 중 normal이 가장 많은 Case에 해당하였기 때문으로 보인다.

하지만 그림 1의 model1의 일반 데이터만을 학습한 모델은 1, 4번의 class 분류에 대해 미미한 오분류가 확인되었고, 그림 2의 2번째 모델인 부정 샘플링을 학습한 모델에서는, 추가된 학습 데이터로 성능이 더 정확해졌음을 확인할 수 있었

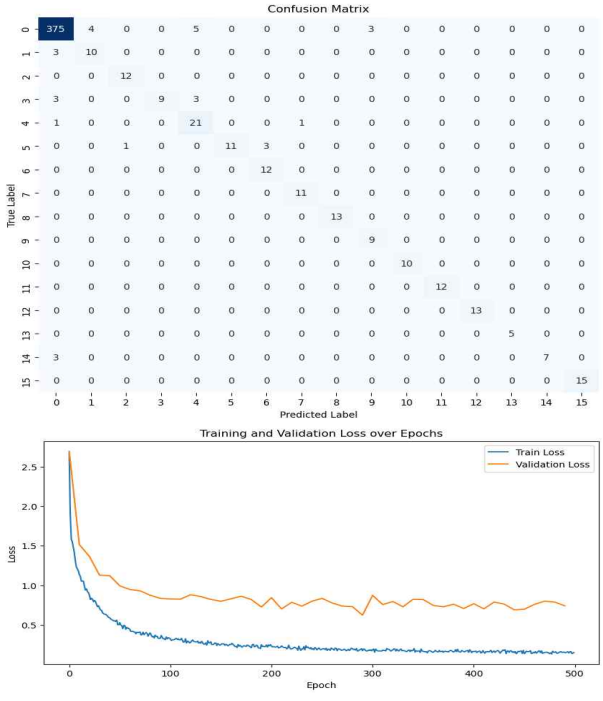

그림 5. GCN모델에 일반 데이터를 학습한 Confusion matrix 및 Loss function 그래프(Model1)

Fig. 5. A confusion matrix and loss function graph train with normal Data with GCN (Model1)

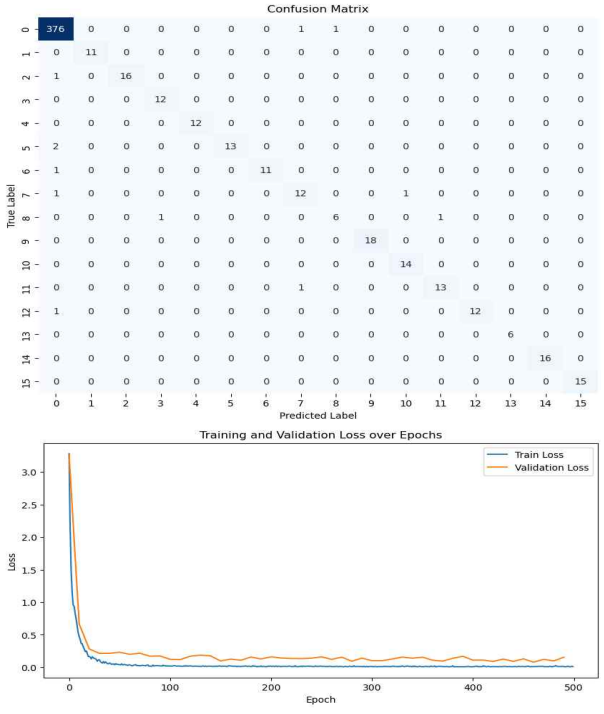

그림 6. GCN모델에 일반 부정 샘플링 데이터를 학습한 Confusion matrix 및 Loss function 그래프(Model2)

Fig. 6. A confusion matrix and loss function graph train with normal negative sampling data with GCN (Model2)

다. 예를 들어, matrix는 1번에 비해 오분류가 줄어들었음을 확인할 수 있었다. 또한, 1번 클래스에 대해서는 2건만이 잘못된 분류를 확인할 수 있었다. 또한 model3의 레이블 교체 기반 부정 샘플링의 경우, 거의 완벽한 분류 정확도를 보이고 있었다.

4-2 Loss Function 분석

Loss Function 또한 Confusion Matrix와 비슷한 양상을 보였다. 먼저 일반 데이터를 학습한 모델은 Epoch가 300에 지점에서 최저점에 도달하였지만, 부정 샘플링을 학습한 나머지 모델들은 전부 50 근처의 Epoch에서 Loss의 수렴에 도달하였다. 또한 실험의 전체 Epoch를 10단위로 조회해본 결과, 일반 데이터를 학습한 Model1은 Epoch 290에서 Loss가 0.1730, Val Loss가 0.6236을 기록한 반면, 일반 부정 샘플링을 학습한 Model2는 Epoch 460에서 Loss가 0.0090, Val Loss가 0.0779을 기록하였고, Model3는 Epoch에서 340, Loss는 0.0108, Val Loss는 0.0953을 기록하였다.

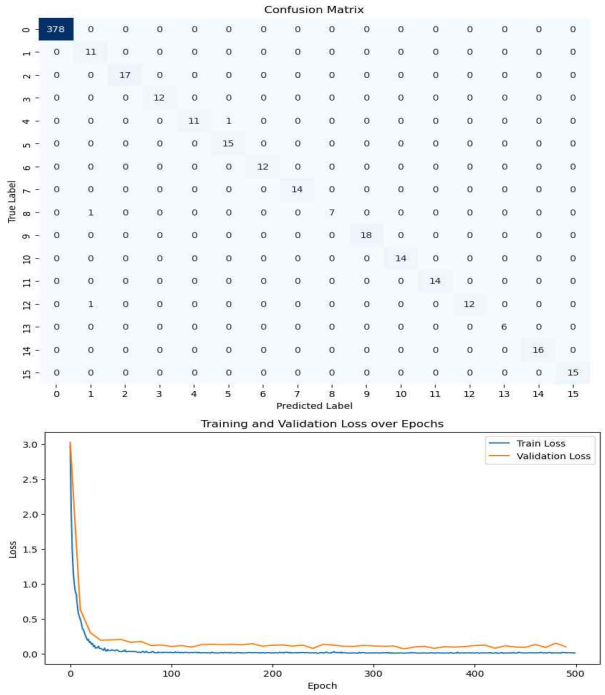

그림 7. GCN모델에 레이블 교체 기반 부정 샘플링 데이터를 학습한 Confusion matrix 및 Loss function 그래프(Model3)

Fig. 7. A confusion matrix and loss function graph train with lable swap based negative sampling data with GCN (Model3)

이 결과는 부정 샘플링을 학습하지 못한 첫 번째 모델과 비교하면, 나머지 두 모델은 확실한 오답의 학습을 통해, 모델이 충분한 오답의 차이를 인식하여 오차율이 현저히 줄어든 것으로 예상된다. 또한 레이블 교체 기반 부정 샘플링 데이터를 학습한 모델의 그래프는 일반 부정 샘플링을 학습한 모델의 그래프에 비해 좀 더 Train loss와 유사한 곡선을 그리는 것으로 보아, 실제 학습한 데이터와 유사하지만, 정답이 아닌 데이터를 학습하는 것이 학습의 효과를 늘리고 있는 것을 확인할 수 있었다.

4-3 평가지표 분석

본 표 1은 모델에 대한 평가지표 데이터를 표로 확인한 값이다. 여기서 Model1은 전처리된 데이터만 학습시킨 모델, Model2는 전처리 데이터와 랜덤 부정 샘플링을 학습시킨 모델, Model3은 전처리 데이터에 레이블 교체 부정 샘플링을 학습시킨 모델이다.

표 1. 전체 모델에 대한 Precision, Recall, F1-Score, Accuracy 지표

Table 1. Precision, recall, F1-Score, and accuracy metrics for the overall model

	Model1	Model2	Model3
Precision	0.9537	0.9792	0.9954
Recall	0.9478	0.9791	0.9948
F1-Score	0.9474	0.9788	0.9948
Accuracy	0.9478	0.9791	0.9948

해당 지표를 봤을 때, 단순한 정확도 성능 지표인 Accuracy 상으로는 Model3가 가장 낮은 지표를 기록한 것으로 확인되지만, 이는 데이터의 내재된 비중이 Normal에 해당되는 항목이 가장 많고 실제 이상분류 상황이 15가지의 케이스인 비해 데이터의 전체 비중에 비한다면 Normal보다 적었기 때문이다. 따라서 이러한 지표의 객관적인 확인을 위해 Precision, Recall, F1-Score를 적용하였다.

Precision은 모델이 예측한 항목들 중 올바르게 예측한 분류항목이고, Recall은 실제 분류항목들 중 모델이 올바르게 예측한 항목, F1-Score은 두 지표에 대한 조화 평균의 값이다. 본 데이터에 의하면, Model1은 Accuracy를 제외한 모든 지표에서 낮은 성능을 보이고 있으며, 이는 평균적인 지표에선 높은 정확도를 보이지만 데이터의 비중이 불균형한 다른 클래스에 대한 분류면에선 낮은 성능을 보이는 것으로 확인되었다. 즉, 부정 샘플링을 학습한 Model2와 Model3는 비록 Accuracy 면에선 낮은 값을 기록하였지만, 반대로 Normal 이외의 다른 분류항목에 대한 정확성은 Model1보다 정확하다고 이해할 수 있다.

또한 Model3은 Model1, 2에 비해 높은 Precision, Recall, F1-Score을 기록하였다. 이는 Model1에 비해 클래스의 불균형 데이터와 편향에 대한 저항성이 높았음을 의미한다. 최종적으로, 본 표를 통해서 확인할 수 있는 사항은 레이블

교체 기반 부정 샘플링 기법을 적용한 데이터 학습 모델이 비록 Accuracy면에선 낮은 성능을 보였지만, 세부적인 분류 능력은 일반 데이터를 학습한 모델은 물론, 랜덤한 부정 샘플링을 학습한 모델보다도 높은 성능을 보였음을 확인할 수 있었다.

V. 결 론

본 연구에서는 "레이블 교체 기반 부정 샘플링"기법을 도입하여, 기존의 네트워크 이상 탐지 분류 모델의 학습을 위한 방법으로써 지식 그래프 학습 모델의 성능을 비교하였다. 이 접근 방법은 기존의 방식인 지식그래프의 노드의 정보를 무작위로 변경 하는 것이 아닌, 레이블의 변경할 값을 편향된 데이터셋을 보완 하도록 교체함으로써, 모델이 원본 데이터와 유사한 부정 샘플링을 학습하게 하여 성능을 향상하게 시키는 것을 목표로 한다. 실험 결과, 이 기법을 적용함으로써 전통적인 부정 샘플링 기법에 비해 더 높은 정확도와 나은 일반화 능력을 보여주었다. 또한, 네트워크 장애 분류와 같은 기존의 시계열 기반 학습 모델로 해결하는 것 이외에도 GCN모델을 도입하여 실험함으로써 해당 모델의 활용성 또한 증명한다. 특히 네트워크와 같은 복잡한 feature에 대한 상관관계를 지닌 데이터에 대해서도 효과적인 성능을 보였다. 결론적으로 이 기법은 복잡한 네트워크 적 구조와 다양한 데이터셋에 적용할 수 있으며, 모델 성능의 강화와 동시에 예측 성능을 높이는 방법으로 확인되었다. 향후 연구를 통해, 좀 더 다양한 모델에 대해서 지식 그래프 학습 개선 실험을 적용하여 이 기법의 범용성과 효과를 더 깊이 탐구하고자 한다.

감사의 글

이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 디지털트윈 기반 네트워크 연구(No. RS-2024-00345030, 디지털트윈 기반 네트워크 장애예방 및 운영관리 자동화 기술 개발) 및 2024년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과입니다.(2021RIS-002)

참고문헌

[1] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus, "Training Convolutional Networks with Noisy Labels," arXiv:1406.2080v4, April 2015. https://doi.org/10.48550/arXiv.1406.2080

[2] X. Zhou, Y. Yi, and G. Jia, "Path-RotatE: Knowledge Graph Embedding by Relational Rotation of Path in Complex Space," in *Proceedings of 2021 IEEE/CIC International Conference on Communications in China (ICCC)*, Xiamen, China, pp. 905-910, July 2021. https://doi.org/10.1109/ICCC52777.2021.9580273

[3] Z. Yang, M. Ding, C. Zhou, H. Yang, J. Zhou, and J. Tang, "Understanding Negative Sampling in Graph Representation Learning," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20)*, Online, pp. 1666-1676, July 2020. https://doi.org/10.1145/3394486.3403218

[4] A. Isah, H. Shin, I. Aliyu, R. M. Sulaiman, and J. Kim, "Graph Neural Network for Digital Twin Network: A Conceptual Framework," in *Proceedings of 2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, Osaka, Japan, pp. 1-5, February 2024. https://doi.org/10.1109/ICAIIC60209.2024.10463455

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321-357, June 2002. https://doi.org/10.1613/jair.953

[6] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*, Toulon, France, April 2017. https://doi.org/10.48550/arXiv.1609.02907

[7] Z. Sun, C. Wang, W. Hu, M. Chen, J. Dai, W. Zhang, and Y. Qu, "Knowledge Graph Alignment Network with Gated Multi-Hop Neighborhood Aggregation," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI-20)*, New York: NY, pp. 222-229, February 2020. https://doi.org/10.1609/aaai.v34i01.5354

[8] D. Shanks, "Non-Linear Transformations of Divergent and Slowly Convergent Sequences," *Journal of Mathematics and Physics*, Vol. 34, No. 1-4, pp. 1-42, April 1955. https://doi.org/10.1002/sapm19553411

[9] A. Clauset, C. Moore, and M. E. J. Newman, "Hierarchical Structure and the Prediction of Missing Links in Networks," *Nature*, Vol. 453, No. 7191, pp. 98-101, May 2008. https://doi.org/10.1038/nature06830

[10] J. Robinson, C.-Y. Chuang, S. Sra, and S. Jegelka, "Contrastive Learning with Hard Negative Samples," arXiv:2010.04592v1, October 2020. https://doi.org/10.48550/arXiv.2010.04592

[11] Barcelona Neural Networking Center. Graph Neural Networking Challenge 2023 [Internet]. Available: https://bnn.upc.edu/challenge/gnnet2023/.

이중훈(JungHoon Lee)

2010년～2017년: 전남대학교 학사과정 졸업(물리)
2023년～현　재: 전남대학교 석사과정 진학(지능전자 컴퓨터공학)

※관심분야 : 지식 그래프, 디지털트윈네트워크 AI

오승민(Seungmin Oh)

2015년～2019년: 한국 나사렛 대학교 학사과정 졸업(디지털콘텐츠)
2019년～2021년: 전남대학교 석사과정 졸업(ICT융합시스템)
2021년～현　재: 전남대학교 박사과정 진학(지능전자 컴퓨터공학)

※관심분야 : 딥러닝, 머신러닝, 컴퓨터 비전, 메타 학습

김광기(Kwangki Kim)

1998년～2002년 : 항공우주대학교 학사과정 졸업(전자엔지니어링)
2002년～2004년 : KAIST 석사과정 졸업(정보통신공학)
2004년～2011년 : KAIST 박사과정 졸업(정보통신공학)
2012년～2013년: 삼성DMC연구소 연구원
2013년～현　재: 한국 나사렛 대학교 부교수

※관심분야 : 음성/음향, 신호처리, 멀티미디어, 디지털 콘텐츠, 회로망

한민수(Min-Soo Hahn)

1976년～1979년: 서울 국립대학교 학사과정 졸업(전자공학)
1980년～1981년: 서울 국립대학교 석사과정 졸업(전자공학)
1982년～1985년: 플로리다 대학(University of Florida) 박사과정 졸업(전자공학)
1982년～1985년: 대전 한국표준과학연구원 소속
1990년～1997년: ETRI 소속
1998년～현　재: KAIST 전자공학과 및 아스타나IT대학교 컴퓨터 및 데이터과학과 교수

※관심분야 : 음성 및 오디오 코딩, 음성 합성, 음성 변환, 노이즈 감소, VoIP

김진술(JinSul Kim)

1997년～2001년: 미국 유타주 솔트레이크시티 유타대학교(University of Utah, Salt Lake City)학사과정 졸업(컴퓨터과학)
2003년～2005년: 카이스트 석사과정 졸업(정보통신학)
2005년～2008년: 카이스트 박사과정 졸업(정보통신학)
2005년～2009년: 한국전자통신연구원 연구위원
2009년～2012년: 한국 나사렛 대학교 부교수
2012년～현　재: 전남대학교 지능전자컴퓨터공학과 교수

※관심분야 : 디지털트윈네트워크, AI, 멀티미디어, 모빌리티