실세계 그래프 데이터에 대한 공정성 분석

신호중 [1], 이연창 [2], 김상욱 [1*]
[1] 한양대학교 컴퓨터소프트웨어학과
[2] 울산과학기술원 인공지능대학원

hojung.shin@agape.hanyang.ac.kr, yeonchang@unist.ac.kr, wook@hanyang.ac.kr

Fairness Analysis on Real-World Graph Data

Hojung Shin[1], Yeon-Chang Lee[2], Sang-Wook Kim[1*]
[1]Dept. of Computer Science, Hanyang University
[2]Graduate School of Artificial Intelligence, UNIST

요 약

그래프 신경망(Graph Neural Network, GNN)은 실세계 그래프 데이터에 대한 다양한 다운스트림 작업들에서 우수한 성능을 보여 왔다. 그러나, 최근 연구는 GNN 의 예측 결과가 데이터 내 특정 집단에 대한 차별을 내포할 수 있음을 지적했다. 이러한 문제를 해결하기 위해, 공정성을 고려할 수 있는 GNN 방법들이 설계되어 오고 있으나, 아직 실세계 그래프 데이터가 공정성 관점에서 어떠한 특성을 가지고 있는지에 대한 분석은 충분히 이루어지지 않았다. 따라서, 본 논문에서는 다양한 공정성 평가 지표를 활용하여 실세계 그래프 데이터의 공정성을 비교 분석한다. 실험 결과, 실세계 그래프 데이터들은 도메인 혹은 평가 지표에 따라 다른 특성을 가진다는 것을 확인하였다.

1. 서론

실세계 다양한 도메인의 데이터는 개체 간 관계를 나타내는 그래프(예: 소셜 네트워크, 단백질 상호 작용 그래프, 도로 교통망)의 형태로 표현될 수 있다. 이러한 그래프 데이터를 효과적으로 분석하고 활용하기 위해, 최근에는 다양한 그래프 신경망(Graph Neural Network, GNN) 방법들이 제안되었다 [3, 4, 5, 6].

그러나, GNN 의 예측 결과는 특정 집단에 대해 차별적일 수 있다 [3, 4, 5, 6]. 이러한 잠재적인 차별은 그래프 내 유사한 특성을 가진 개체들 간의 관계 형성이 서로 다른 특성을 가진 개체들 간의 관계 형성보다 빈번하게 발생함에 따라 야기된다 [1, 2]. 더 나아가, GNN 의 이웃 정보 전파 메커니즘은 이러한 그래프 데이터의 동종 선호(homophily) 특성을 심화하여 데이터 편향을 악화시킬 수 있다 [3, 5, 6].

이를 해결하기 위해, 최근에는 공정성을 고려할 수 있는 다양한 GNN 방법들이 제안되어 오고 있다 [3, 4, 5]. FairGNN [3]은 적대적 신경망 구조를 기반으로 최종 노드 임베딩이 해당 노드의 소속 집단을 예측하기 어렵도록 학습한다. EDITS [4]는 특성 편향(attribute bias)와 구조 편향(structure bias)을 감소시킬 수 있는 GNN 방법을 제안한다. FairVGNN [5]은 GNN 의 이웃 정보 전파 메커니즘에서 민감 특성 누수를 최소화할 수 있는 방법을 제안한다.

그러나, 이러한 연구들은 실세계 그래프 데이터가 공정성 관점에서 어떠한 특성을 가지고 있는지에 대한 분석을 충분히 수행하지 않았다. 따라서, 본 논문에서는 다양한 공정성 평가 지표를 기반으로 실세계 그래프 데이터의 공정성을 면밀히 분석해보고자 한다.

2. 공정성 평가 지표

본 논문에서, 우리는 네 가지 공정성 평가 지표를 활용한다 [3, 6]. 이러한 지표들에서, 공통적으로 M은 노드 임베딩의 총 차원수, W는 Wasserstain-1 거리, pdf는 확률밀도함수, 그리고 s_u는 노드 u가 속한 집단을 나타낸다. G와 E는 각각 그래프와 간선 집합을 나타낸다. 또한, 우리는 데이터 내 개체들에 대한 집단은 두 가지만 존재(예: 남/여)한다고 가정한다.

2.1 임베딩 분포 측면의 공정성 평가 지표

집단 간 노드 임베딩의 분포 차이가 클수록, GNN 방법들은 특정 노드의 소속 집단을 알아내기 쉬울 것이므로 [4], 우리는 임베딩 분포 측면의 두 가지 공정성 평가 지표를 활용한다.

특성 편향 (Attribute Bias). 특성 편향 b_{attr}은 노드들의 특성 행렬(attribute matrix) X에서 집단 간의 분포 차이를 나타내며, 다음과 같이 계산된다 [4].

$$b_{attr} = \frac{1}{M}\sum_{m} W\left(pdf(X_m^0), pdf(X_m^1)\right), \qquad (1)$$

여기서, X_m^i은 i번째 집단에 속한 노드들의 m번째 특성 차원 값들을 나타낸다.

구조 편향(Structure Bias). 구조 편향 b_{stru}은 주어진 그래프 데이터에서 GNN 방법을 수행한 이후에 학습된 노드 임베딩들에서 집단 간 분포 차이를 나타내며, 다음과 같이 계산된다 [4].

$$b_{stru} = \frac{1}{M}\sum_{m} W\left(pdf(R_m^0), pdf(R_m^1)\right), \quad (2)$$

여기서, R_m^i은 GNN 방법이 수행된 이후 i번째 집단에 속한 노드들의 m번째 임베딩 차원 값들을 나타낸다.

2.2 이웃 분포 측면의 공정성 평가 지표

그래프 데이터에서 집단 내 연결이 많고 집단 간 연결이 적다면, GNN 방법들은 집단 내 임베딩은 유사하게, 집단 간 임베딩은 유사하지 않도록 학습을 수행한다. 따라서, 우리는 그래프 데이터 내 이웃 분포 측면의 두 가지 공정성 평가 지표를 활용한다.

동종 선호 비율 (Homophily Ratio). 동종 선호 비율 $h(G)$은 주어진 그래프 데이터 G에서 집단 내 간선의 상대적 비율을 나타내며, 다음과 같이 계산된다 [6].

$$h(G) = \frac{\sum_i C_{ii}}{\sum_i \sum_j C_{ij}} = \frac{\sum_i C_{ii}}{|E|}, \quad (3)$$

$$C_{ij} = |\{(u,v): (u,v) \in E \wedge s_u = i \wedge s_v = j\}|, \quad (4)$$

여기서, C_{ij}는 집단 i와 j간의 간선의 빈도수를 나타낸다. $h(G)$는 1 에 가까울수록 같은 집단 내 노드들 간의 간선이 많다는 것을 의미하며, 0 에 가까울수록 다른 집단 간 노드들의 간선이 많다는 것을 의미한다.

이웃 공정성 (Neighborhood Fairness). 이웃 공정성 $\mathbb{F}(\mathrm{G})$은 주어진 그래프 데이터 G에서 이웃들의 평균 엔트로피를 나타내며, 다음과 같이 계산된다 [6].

$$\mathbb{F}(\mathrm{G}) = \frac{1}{|V|}\sum_{i \in V} \mathbb{F}(\mathbf{p}_i) \quad (5)$$

$$\mathbb{F}(\mathbf{p}_i) = -\sum_{k} p_{ik} \log p_{ik}, \quad (6)$$

여기서, p_{ik}는 노드 i의 이웃들의 소속 집단이 k일 확률을 나타내며, $\mathbb{F}(\mathbf{p}_i)$는 임의의 노드 i의 엔트로피를 나타낸다. 결과적으로, $\mathbb{F}(\mathrm{G})$는 값이 클수록 그래프의 이웃 분포가 균일하다는 것을 의미하며, 작을수록 특정 집단에 편향되었다는 것을 의미한다.

3. 실험

<표 1> 데이터 통계

데이터 집합	노드	간선	집단 내 간선	집단 간 간선
Pokec_z	67,797	651,856	621,337	30,519
Pokec_n	66,569	550,331	526,038	24,293
NBA	403	10,822	7,887	2,935
German	1,000	22,242	17,998	4,244
Credit	30,000	1,436,858	1,379,322	57,536
Recidivism	18,876	321,308	172,259	149,049

본 논문에서는 6 가지의 실세계 그래프 데이터를 사용한다: Pokec_z, Pokec_n, NBA, German, Credit, Recidivism. <표 1>은 본 논문에서 사용한 데이터 집합들의 통계를 보여준다.

<표 2>는 각 데이터에 대해 공정성 평가 지표를 측정한 결과를 보여준다. 여기서, 특성 편향과 구조 편향은 값이 클수록, 동종 선호 비율은 1 에 가까울수록, 이웃 공정성은 0 에 가까울수록 주어진 그래프 데이터에 높은 편향이 내재되어 있음을 의미한다.

실험 결과, 우리는 모든 데이터 집합들에서 구조 편향이 특성 편향보다 크게 나타나타는 것을 확인하였다. 이러한 결과는 GNN 방법이 집단 간의 분포 차이를 증폭시킨다는 것을 보여준다. 또한, 동종 선호 비율과 이웃 공정성 관점의 편향이 큰 데이터 집합들(예: Pokec_z)에서 구조 편향과 특성 편향 간의 차이가 더 크게 나타났다. 이러한 결과는 집단 내 간선과 집단 간 간선의 비율 차이가 GNN 방법이 수행된 후의 편향 심화 정도에 영향을 미침을 의미한다.

<표 2> 데이터 별 공정성 평가 지표 측정 결과

데이터 집합	특성 편향	구조 편향	동종 선호 비율	이웃 공정성
Pokec_z	0.009	0.179	0.953	0.132
Pokec_n	0.142	0.248	0.956	0.114
NBA	4.148	5.898	0.729	0.499
German	6.328	10.396	0.809	0.418
Credit	2.463	4.451	0.960	0.158
Recidivism	0.953	1.098	0.536	0.657

4. 결론

본 논문에서는 실세계 그래프 데이터를 다양한 공정성 평가 지표를 통해 분석하였다. 실험 결과, GNN 방법이 데이터의 편향을 증폭시킬 수 있으며, 이는 집단 내/외 간선 비율 차이와 관련이 있음을 확인하였다. 이러한 관찰은 해당 데이터를 활용하는 연구자들에게 유용한 정보를 제공할 것으로 기대된다.

사사

이 논문은 2023 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2022-00155586, 실세계의 다양한 다운스트림 태스크를 위한 고성능 빅 하이퍼그래프 마이닝 플랫폼 개발(SW 스타랩), No. 2020-0-01373, 인공지능대학원지원(한양대학교), No.2018R1A5A7059549).

참고문헌

[1] Yuxiao Dong et al., Do the Young Live in a “Smaller World” Than the Old? Age-Specific Degrees of Separation in a Large-Scale Mobile Communication Network, arXiv preprint arXiv:1606.07556, 2016.

[2] Tahleen A Rahman et al., Fairwalk: Towards Fair Graph Embedding, IJCAI, 2019.

[3] Enyan Dai et al., Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information, WSDM, 2021.

[4] Yushun Dong et al., Edits: Modeling and mitigating data bias for graph neural networks, The Web Conference, 2022.

[5] Wang et al., Improving fairness in graph neural networks via mitigating sensitive attribute leakage, KDD, 2022.

[6] A. Chen et al., Fairness-aware graph neural networks: A survey, arXiv preprint arXiv:2307.03929, 2023.