카운팅 블룸 시그니처 기반 구조 표현을 활용한 그래프 신경망 링크 예측 기법

강지수, *변하영
명지대학교 전자공학전공

toujours89@mju.ac.kr, *hbyun@mju.ac.kr

Graph Neural Network-Based Link Prediction Using Structural Representations with Counting Bloom Signatures

Jisoo Kang, *Hayoung Byun
Myongji University

요 약

본 연구에서는 블룸 시그니처(Bloom Signature) 기반 링크 예측 방법을 확장하여, 2-비트 카운팅 블룸필터를 이용해 그래프 구조를 표현하고, 카운팅 정보를 활용해 이웃 구조 간 중첩 정도를 반영할 수 있는 구조적 피처를 생성한다. 기존 방법에서 사용되던 구조적 피처를 유지한 상태에서, 카운팅 기반 피처를 추가로 도입한다. 이러한 구조적 피처는 그래프 신경망을 통해 학습된 노드 임베딩과 결합되어 링크 예측기에 입력된다. 실제 그래프 데이터셋을 대상으로 실험한 결과, 제안하는 방법은 기존 방법과 동일한 메모리를 유지하면서도, 향상된 링크 예측 성능을 보였다.

Ⅰ. 서론

그래프 기반 데이터에서 링크 예측(link prediction)은 아직 연결되지 않은 두 노드의 연결 가능성을 예측하는 핵심 문제로, 추천 시스템, 지식 그래프, 네트워크 분석 등 다양한 응용 분야에서 중요하게 다루어진다. 최근에는 그래프 신경망(Graph Neural Network, GNN)을 활용한 링크 예측 연구와 함께, 구조적 피처(Structural features)를 결합하는 연구가 진행되고 있다. 그중 블룸 시그니처(Bloom signature)를 활용한 링크 예측 연구는 그래프의 이웃 구조를 해시 기반의 압축된 표현으로 인코딩한다. 이 표현으로부터 다양한 구조적 휴리스틱을 근사적으로 도출할 수 있으며, 이를 GNN 기반 링크 예측기에 입력으로 활용할 수 있음을 보였다[1].

본 연구에서는 이러한 표현 방식을 확장하여 이웃 구조의 중첩 정도에 대한 정보를 함께 반영하기 위해, 기존 블룸 시그니처를 2 비트 카운팅 블룸 필터(Counting Bloom Filter, CBF)로 대체하고, 카운팅 기반 피처를 추가하였다. 실제 데이터셋을 대상으로 실험한 결과, 기존 연구와 동일한 메모리 예산 하에서 기존 방법 대비 향상된 링크 예측 성능을 보인다.

Ⅱ. 본론

기존의 GNN 과 블룸 시그니처 기반 연구에서는 블룸 시그니처를 사용하여 그래프의 이웃 구조를 해시 기반의 이진 벡터로 표현하였다[1]. 이 방식은 인접 행렬이나 이웃 집합을 직접 저장하지 않고도, 시그니처 간 단순한 비트 연산을 통해 이웃 구조의 교집합(intersection), 여집합(complement), 포함도(containment), 코사인 유사도(cosine similarity) 등 다양한 구조적 휴리스틱을 추정할 수 있다. 이 구조적 휴리스틱 피처는 링크 예측을 위한 피처 $\hat{S}(u,v)$로 활용한다.

본 연구에서는 이러한 블룸 시그니처 기반 접근을 확장하여, 이진 값 대신 카운팅 정보를 저장하는 CBF를 구조적 표현으로 사용한다. CBF는 동일한 해시함수를 사용하여 두 노드가 기여한 이웃 수를 저장함으로써, 이웃 구조 간 중첩 정도를 직접적으로 표현할 수 있다.

그림 1은 노드 쌍 (u, v)의 링크를 예측하기 위하여 두 노드의 k -hop 이웃 구조가 각각 $CBF_u^{(j)}, CBF_v^{(j)}$ 로 표현되는 과정이다 (여기서 $j = 1, \ldots, k$ 이며 $k = 2$ 로 가정한다). 이때 각 이웃 노드는 하나의 해시 함수를 통해 CBF의 해시 공간에 매핑되며, 동일한 해시 위치에 매핑된 이웃 노드들이 누적되면서 두 노드가 공통으로 기여한 구조적 중첩을 카운트로 표현할 수 있다.

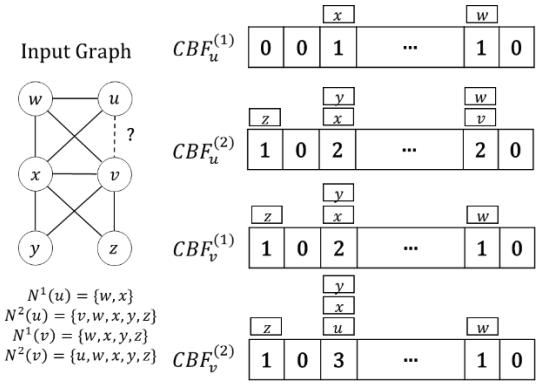

그림 1. CBF 에서 이웃 구조 중첩 표현 과정

이러한 중첩 정보는 노드 쌍의 구조적 피처로 활용된다. 즉, 이웃 구조의 중첩 정도를 근사한 구조적 피처 $\hat{S}_{\mathrm{CBF}}(u,v)$ 를 생성하며, 식 (1)은 $\hat{S}_{\mathrm{CBF}}(u,v)$ 의 각 성분에 대해 나타낸 것이다($j = 1, \ldots, k$). $CBF_u^{(j)}, CBF_v^{(j)}$의 동일한 해시 위치 i에서 두 노드의 카운트 값 $CBF_u^{(j)}[i]$와 $CBF_v^{(j)}[i]$ 중 더 작은 값을 취함으로써 두 노드 쌍의

교집합을 근사하며, CBF의 사이즈 m 에 대해 최소 카운트 값을 집계함으로써 $\hat{S}_{\mathrm{CBF}}(u,v)$를 생성한다.

$$\hat{S}^{(j)}{}_{CBF}(u,v) = \sum_{i=0}^{m-1} \min\left(CBF_u^{(j)}[i], CBF_v^{(j)}[i]\right) \quad (1)$$

그림 2는 노드 쌍 $(u,\ v)$에 대한 링크 존재 확률 $\hat{A}_{uv}$를 얻기 위해, GNN으로부터 얻은 노드 임베딩과 구조적 엣지 피처를 결합하여 활용하는, 제안하는 링크 예측 기법을 나타낸다. 입력 그래프에서 각 노드는 GNN의 메시지 패싱 과정을 통해, 자신의 이웃 정보를 반복적으로 집계하여 이웃 구조를 노드 수준에서 학습하며, 이 과정의 최종 출력으로 노드의 히든 상태(hidden state)인 노드 임베딩 $h_u,\ h_v$를 얻는다. 링크 예측 단계에서는 두 노드 임베딩의 상호작용을 반영하기 위해 원소별 곱 $h_u \circ h_v$을 계산하고, 이를 링크 예측기의 기본 입력으로 사용한다.

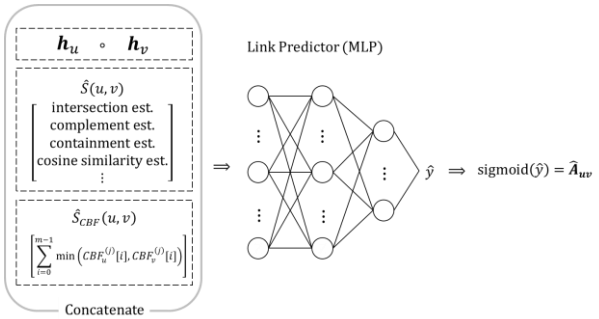

그림 2. 제안하는 GNN 기반 링크 예측 기법

기본적인 GNN 기반 링크 예측은 두 노드 임베딩만을 사용하여 식 (2)과 같이 확률 $\hat{A}_{uv}$를 얻는다. 그러나 이 표현만으로는 이웃 구조 관계를 직접적으로 반영하는 데 한계가 있다. 이를 보완하기 위해 기존 블룸 시그니처 기반 방법은 노드 쌍의 구조적 관계를 요약한 구조 피처 $\hat{S}(u,v)$를 노드 임베딩과 함께 결합하여 링크 예측에 활용한다.

$$\hat{A}_{uv} = \mathbf{sigmoid}(\mathbf{MLP}\ (h_u \circ h_v)) \quad (2)$$

본 연구는 CBF를 블룸 시그니처를 포함하는 일반화된 표현으로 사용하여, 기존 블룸 시그니처에서 사용되던 구조 피처 $\hat{S}(u,v)$를 동일한 방식으로 계산할 수 있도록 한다. 이러한 기존 피처 구성을 유지한 상태에서, 카운팅 정보를 활용한 추가적인 구조 피처 $\hat{S}_{CBF}(u,v)$를 함께 도입한다. 최종적으로, 제안하는 방식의 링크 예측기는 GNN으로부터 얻은 $h_u \circ h_v$ 와 구조적 피처 $\hat{S}(u,v)$, $\hat{S}_{CBF}(u,v)$를 연결(concatenation)한 벡터를 입력 받아, 식 (3)과 같이 링크 예측 확률 $\hat{A}_{uv}$을 출력한다. 이러한 구성은 GNN 백본의 메시지 패싱 구조를 그대로 유지하면서, 링크 예측 단계에서만 구조적 정보를 추가로 활용함으로써 예측 성능 향상에 기여한다.

$$\hat{A}_{uv} = \mathbf{sigmoid}(\mathbf{MLP}(h_u \circ h_v)||\hat{S}(u,v)||\hat{S}_{CBF}(u,v)) \quad (3)$$

Ⅲ. 성능 평가

CBF 기반 구조적 피처의 성능 향상을 검증하기 위해, 서로 다른 특성을 가진 실제 그래프 데이터 셋을 대상으로 링크 예측 성능을 평가하였다[2]. 실험은 Open Graph Benchmark (OGB) 데이터셋 중, 밀도가 높은 ddi와 희소한 collab 그래프를 대상으로 수행하였다. 모든 실험에서 GNN 백본 모델과 학습 하이퍼파라미터는 블룸 시그니처의 최적의 케이스로 세팅하였으며, 제안하는 방식에서는 동일한 메모리 하에서 실험하기 위해 2 비트 CBF의 사이즈 m을 블룸 시그니처의 $\frac{1}{2}$로 조정하였다.

표 1은 각 데이터에 대해 6번의 시행을 반복할 때, 블룸 시그니처와 CBF 방식에 대한 링크 예측 결과이다. Hits@K는 모델이 예측한 상위 K개 후보 안에 정답이 포함되는 비율을 나타낸다. 두 데이터셋 모두에서 CBF를 사용한 경우가 더 높은 링크 예측 성능을 보였으며, 특히 상대적으로 밀도가 높은 ddi 그래프에서는 뚜렷한 성능 개선이 나타난다. 이는 구조적 중첩 정도를 반영한 카운팅 기반 피처를 추가함으로써, 링크 예측 문제에 유용한 구조 정보를 제공할 수 있음을 시사한다. 따라서, 실험 결과를 통해 제안하는 방식이 구조 정보를 효율적으로 표현하며, 메모리 효율성과 표현력 측면에서 기존 방법 대비 효과적임을 보여준다.

표 1. 구조 표현 방식에 따른 링크 예측 성능 비교 (%)

Dataset	Metric	Performance	
		Bloom Signature	CBF
ddi	Hits@20	59.38 ± 10.71	66.47 ± 9.43
collab	Hits@50	61.24 ± 0.58	63.22 ± 0.91

Ⅳ. 결론

본 연구에서는 그래프 링크 예측 문제에서 구조적 표현 방식의 중요성에 주목하고, 기존 블룸 시그니처 기반 접근을 확장하여 CBF를 활용한 구조적 피처 생성 방법을 제안하였다. 제안한 방법은 카운팅을 통해 이웃 구조 중첩의 강도를 반영하며, 동일한 메모리를 유지한 조건에서 구조적 정보를 보다 풍부하게 표현할 수 있다. 실제 그래프 데이터 셋에서 CBF 기반 구조 표현이 블룸 시그니처 대비 향상된 링크 예측 성능을 보였다.

향후 연구에서는 다양한 그래프 데이터셋에 대한 적용 가능성을 검토하고, 카운팅 범위 및 구조적 피처 추가 설계에 따른 성능 변화를 체계적으로 분석하고자 한다.

ACKNOWLEDGMENT

본 과제(결과물)는 2025 년도 교육부 및 경기도의 재원으로 경기 RISE 센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다. (2025-RISE-09-A15)

참 고 문 헌

[1] Tianyi Zhang, Haoteng Yin, Rongzhe Wei, Pan Li, and Anshumali Shrivastava, “Learning Scalable Structural Representations for Link Prediction with Bloom Signatures,” *Proceedings of the ACM Web Conference (WWW), 2024*, pp. 980–991.

[2] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec, “Open graph benchmark: Datasets for machine learning on graphs,” *Advances in Neural Information Processing Systems,* vol. 33, 2020, pp. 22118–22133.