ISSN 2383-6318(Print) / ISSN 2383-6326(Online)
KIISE Transactions on Computing Practices, Vol. 30, No. 10, pp. 543-547, 2024. 10
https://doi.org/10.5626/KTCP.2024.30.10.543

자기지도학습기반 그래프 임베딩을 위한 네거티브 샘플링 알고리즘

(Negative Sampling Strategy for Self-Supervised Graph Embedding)

연 정 흔 † 신 현 정 ††
(JeongHeun Yeon) (Hyunjung Shin)

요 약 최근 그래프기반 태스크에서 우수한 성능을 보이는 접근법은 대부분 지도학습기반의 그래프신경망으로 학습 시 라벨 정보는 필수적이다. 하지만 라벨링 작업은 매우 지루하고 많은 비용이 필요한 작업이다. 따라서 최근 라벨을 활용하지 않고 데이터 간의 관계 정보만을 활용해 지도학습 태스크를 수행할 수 있는 자기지도학습 방법에 대한 연구가 대두되고 있다. 그 중 Graphical Mutual Information은 그래프 데이터에 적용할 수 있는 자기지도학습 방법으로 각 노드를 특징지어 임베딩하기에 적절하다. 하지만 직접적 연결정보만을 활용하는 특성상 그래프의 전역적인 군집구조를 학습하는데 한계가 있다. 따라서 본 연구에서는 그래프 군집구조를 학습에 활용할 수 있는 마지널 노드 인디케이터 방법을 제안한다. 벤치마크 데이터셋에 대한 비교실험을 통해 기존 GMI 대비 성능이 개선될 수 있음을 검증하였다.

키워드: 그래프 표현 학습, 자기지도학습, 네거티브 샘플링, 그래프 기반 상호정보량, 그래프 대조 학습

Abstract In recent tasks involving graphs, the prevailing approaches mostly rely on supervised learning-based graph neural networks. These approaches require label information during training, which can be both tedious and expensive to obtain. As a result, there is a growing interest in self-supervised learning methods that use only relational data between elements, eliminating the need for labels. One such method is Graphical Mutual Information (GMI), which has proven effective in embedding each node based on its characteristics. However, GMI has a limitation: it solely relies on direct connection information, which hinders its ability to learn the global cluster structure of graphs. To overcome this limitation, our research proposes the Marginal Node Indicator method, which leverages the graph's cluster structure in the learning process. Through comparative experiments on benchmark datasets, we have validated that this approach improves performance compared to existing GMI methods.

Keywords: graph representation learning, self-supervised learning, negative sampling, graph-based mutual information, graph contrastive learning

· 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2022-0-00653, 보이스피싱 정보 수집·가공 및 빅데이터 기반 수사지원시스템 개발). 이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2021R1A2C2003474, 치매 정밀의료 및 진단 다각화를 위한 인공지능 모델 개발). 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 인공지능융합혁신인재양성사업 연구 결과로 수행되었음(IITP-2024-No.RS-2023-00255968). 본 연구는 한국연구재단의 BK21 4단계사업 연구 결과로 수행되었음(NRF5199991014091)
· 이 논문은 2023 한국소프트웨어종합학술대회에서 '자기지도학습기반 그래프 임베딩을 위한 네거티브 샘플링 알고리즘'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 아주대학교 인공지능학과 학생
yjh970@ajou.ac.kr
†† 종신회원 : 아주대학교 산업공학과 교수(Ajou Univ.)
shin@ajou.ac.kr
(Corresponding author임)

논문접수 : 2024년 3월 20일
(Received 20 March 2024)
논문수정 : 2024년 5월 15일
(Revised 15 May 2024)
심사완료 : 2024년 5월 22일
(Accepted 22 May 2024)

정보과학회 컴퓨팅의 실제 논문지 제30권 제10호(2024. 10)

1. 서 론

그래프 데이터에서 유의미한 정보를 추출하기 위해선 그래프의 구조를 명확히 파악하는 것이 중요하다. 데이터를 그래프로 표현하면 데이터 간의 관계 정보를 알 수 있기 때문에 이를 활용한 노드 분류, 연결성 예측, 커뮤니티 탐지 등의 태스크를 수행할 수 있다. 최근 지도학습기반의 GNNs을 활용한 접근법이 다양한 그래프 기반 태스크에서 우수한 성능을 보이고 있다. 하지만 모델 학습 시 정답이 되는 라벨이 반드시 필요하고, 현실에서 라벨을 얻는 작업은 매우 지루하고 많은 비용을 요구한다. 따라서 최근 라벨을 활용하지 않고 그래프 내의 정보만을 활용하는 자기지도학습 방법이 활발히 연구되고 있다[1]. 이 방법은 라벨 정보 없이 데이터 속성 정보와 데이터 간의 관계정보만으로 사전학습하여 기존 지도학습 태스크를 수행할 수 있으므로 매우 효율적이다. 그 중 GMI(Graphical Mutual Information)[2]는 그래프에 적용할 수 있는 자기지도학습 방법론으로, 대조 학습을 기반으로 노드의 속성 정보와 그래프의 구조(연결성)정보만을 학습에 활용한다. 이 방법은 노드 임베딩 시 positive/negative sample을 정의해 target node와 positive sample간에는 유사하게 임베딩함과 동시에 negative sample과는 대조되게 임베딩되도록 학습시킨다. GMI에서는 각 target node마다 positive sample을 k-hop 이내에 직접적으로 연결이 있는 이웃 노드로 정의하고, 이를 제외한 나머지 노드를 negative sample로 정의한다. 따라서 노드는 직접적으로 관계가 있는 주변 정보만을 포함하게 되므로 각 노드를 표현하기에 적절하다.

그래프 데이터는 정의상 서로 관련 있는 데이터 간에 연결이 있으므로 서로 유사한 노드끼리(동일한 레이블) 군집을 형성하는 동형(homophily) 특성을 갖을 수 있다.

이 때, 군집 간의 경계를 의미하는 마진(margin)에 있는 노드의 경우 직접적으로 연결된 이웃 노드의 레이블이 혼재된 경우가 많다. GMI는 직접적 주변정보만을 활용하므로 서로 다른 군집에 있으나 공통된 이웃을 공유하는 노드를 구분할 수 없게끔 학습하게 된다(그림 1). 이는 서로 다른 군집에 속한다는 구조정보를 반영하지 못하므로 다운스트림 태스크에서 성능 저하를 일으킬 수 있다. 하지만 그림 1을 보면 알 수 있듯, 그래프 전체 군집구조를 알고 있다면 두 노드의 차이를 쉽게 구별할 수 있다. 따라서 본 연구는 기존 GMI 방법론에서 샘플링 시 그래프의 군집정보를 추가로 반영한 positive/negative sampling 전략을 제안한다. 이를 통해 대조 학습을 활용한 그래프 임베딩 시, 마지널 노드(marginal node)에 더 강건한 방법론을 제안하며, 그래프 벤치마크 데이터셋에 대한 비교실험을 통해 성능 개선을 검증하였다.

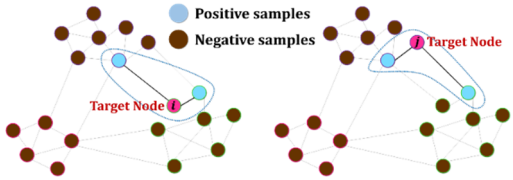

그림 1 네트워크 마진에서 GMI의 ositive/negative 샘플 정의

Fig. 1 Definition of Positive/Negative Samples for GMI at the Network Margin

2. 제안 방법론

그래프의 군집구조를 활용하기 위해 그래프 클러스터링이 선행된다. 본 연구에서는 그래프의 연결정보로 군집화하는 스펙트럴 클러스터링 알고리즘을 활용했지만 이는 각 노드의 군집 라벨이 필요한 것이므로 그래프 파티셔닝과 같은 다른 알고리즘으로 대체 가능하다.

2.1 코어 & 마지널 노드

제안 방법론에서는 그래프의 군집정보로부터 코어 노드와 마지널 노드를 정의한다. 그림 2에서 볼 수 있듯이, 각 군집의 중심부에 위치해 다른 군집과 이웃하지 않는 노드를 코어 노드($Core$), 반대로 군집 간의 경계에 위치해 서로 다른 군집의 노드와 이웃하고 있는 노드를 마지널 노드(M)로 정의한다. C_k는 클러스터의 인덱스이고, $K \in 1, \ldots, k$는 클러스터 수이다.

2.2 이웃기반 상호정보량

그래프에서 앞서 정의한 코어/마지널 노드를 찾기 위해 각 노드 별로 특정 군집에 속한 정도를 정량적으로 측정할 수 있는 Neighbor Mutual Information(NMI)을 제안한다. NMI는 두 변수간 연관성을 측정하기 위한 점별 상호정보량(PMI)을 그래프 공간으로 확장한 개념이다. PMI와 NMI의 정의는 다음과 같다.

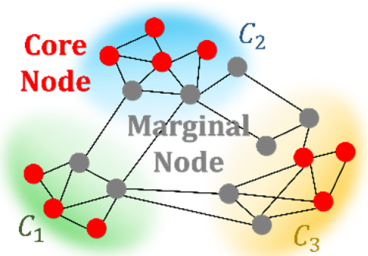

그림 2 코어/마지널 노드

Fig. 2 Core/Marginal Node

$$PMI(X,Y) = log\frac{p(x,y)}{p(x)p(y)} \quad (1)$$

$$NMI_{i,k} := \frac{\frac{|\mathcal{N}(i)\cap C_k|}{N}}{\frac{|\mathcal{N}(i)|}{N}\frac{|C_k|}{N}} = \frac{N\cdot|\mathcal{N}(i)\cap C_k|}{|\mathcal{N}(i)|\cdot|C_k|} \quad (2)$$

$\mathcal{N}(\cdot)$는 노드 i의 이웃노드집합이고, N은 전체 노드 수이다. PMI는 두 확률변수 X, Y 간의 연관성을 동시 발생확률과 각 분포 간의 차이로 측정한다. 유사하게 NMI는 노드와 각 클러스터 간의 연관성을 타겟노드의 이웃노드 중 특정 클러스터에 속하는 노드 수와 (이웃노드 수 × 클러스터에 속한 노드 수) 간의 차이를 측정한다. 즉, 이웃노드와 특정 클러스터 간의 교집합이 클수록 타겟노드와 해당 클러스터 간의 연관성이 큼을 의미한다. 최종적으로 다음과 같이 범위를 [0,1]로 정규화하여 활용한다.

$$Normalised\ NMI_{i,k} \leftarrow \frac{NMI_{i,k}}{\sum_{k=1}^{K} NMI_{i,k}} \quad (3)$$

정규화된 NMI를 통해 모든 노드와 클러스터 간의 연관성을 측정하면 그림 3과 같이 코어 노드와 마지널 노드의 차이를 확인할 수 있다. 코어 노드인 노드 i의 경우, 이웃노드가 전부 C_3에 속하므로 C_3와의 연관성은 매우 높지만 그 외 클러스터와의 연관성은 0이다. 반면 마지널 노드인 노드 j의 경우, 이웃노드의 클러스터가 혼재되어 있어 여러 클러스터와의 NMI 값이 0.5 근처를 갖는다. 따라서 본 연구에서는 NMI의 이러한 특성을 활용해 그래프에서 코어 노드와 마지널 노드를 정량적으로 찾을 수 있는 마지널 노드 인디케이터(Marginal Node Indicator)를 다음과 같이 정의한다.

$$Core(i) := \left\{x_j | NMI_{j,k} \geq \delta \cdot max\left(NMI_{j,k}\right),\ k = c(i)\right\} \quad (4)$$

$$M(i) := \left\{\begin{matrix} x_j | (1-\lambda)\cdot\delta < NMI_{j,k} < \lambda\cdot\delta, \\ i \neq j, c(j) \neq c(i) \end{matrix}\right\} \quad (5)$$

$c(i)$는 노드 i의 클러스터 인덱스를 의미한다. δ는 코어 노드와 마지널 노드를 구분하는 상한을 조절해 코어의 범위를 조절할 수 있고, λ는 마지널의 범위를 조절할 수 있는 하이퍼파라미터이다. 특히, 코어 노드는 타겟

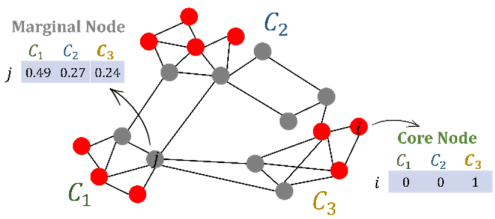

그림 3 노드 i, j와 각 클러스터와의 NMI
Fig. 3 NMI values between node i, j and each cluster

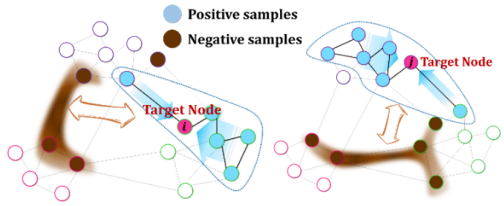

그림 4 제안방법론으로 선정한 positive/negative 샘플 예시
Fig. 4 Example of positive/negative samples selected using the proposed method

노드 i와 동일한 클러스터에 속하는 노드로, 반대로 마지널 노드는 같은 클러스터에 속하지 않은 노드로 정의한다.

2.3 마지널 노드 인디케이터 기반 샘플링 전략

최종적으로 앞서 정의한 마지널 노드 인디케이터를 통해 그래프로부터 각 노드별 코어 노드와 마지널 노드를 찾고, 이를 활용한 positive/negative sampling 전략을 제안한다.

$$\boldsymbol{H} := \{Core(i) \cup \mathcal{N}(i)\} \quad (6)$$

$$\widetilde{\boldsymbol{H}} := M(i) \quad (7)$$

$\boldsymbol{H}$는 positive sample을, $\widetilde{\boldsymbol{H}}$는 negative sample을 나타낸다. 타겟 노드와 유사하게 임베딩할 positive sample 집합을 기존 직접적 이웃노드 $\mathcal{N}(i)$에서 같은 클러스터에 속한 코어 노드를 추가함으로써 각 노드는 클러스터별로 더 유사하게 임베딩되고, negative sample을 타겟 노드가 속하지 않은 클러스터의 마지널 노드로 제한함으로써 그래프의 전체 군집 구조를 반영할 수 있다. 최종적으로 그림 1에서의 노드 i, j는 그림 4과 같이 샘플링 집합이 변하게 된다.

3. 실 험

실험은 총 4개의 벤치마크 데이터셋에 대해 비교실험을 수행하였다. 모두 학습용 그래프와 테스트용 그래프가 동일한 transductive 방식으로 노드 분류 태스크를 수행하였다. 성능은 클래스별 정확도(accuracy) 평균을 사용하였다. 실험에 사용된 데이터의 세부사항은 표 1에 나타내었다. 비교모델로는 학습 시 라벨을 활용하는 지도학습 기반의 모델들과 자기지도학습 방법을 활용하는 모델들을 선정하였다.

3.1 실험

실험은 인용 네트워크로 구성된 벤치마크 데이터셋 4가지에 대해 수행하였다. 각 네트워크의 세부사항은 표 1에서 확인할 수 있으며, 비교모델들과의 그래프 표현 성

표 1 데이터셋 설명
Table 1 Description of Datasets

Task	Dataset	Type	#Nodes	#Edges	#Features	#Classes
Node classification / Transductive	Cora	Citation Network	2,708	5,429	1,433	7
	Citeseer	Citation Network	3,327	4,732	3,703	6
	Chameleon	Citation Network	2,277	31,421	2,325	5
	Squirrel	Citation Network	5,201	198,493	2,089	5

능을 비교하기 위해 지도학습에서의 노드 분류 태스크를 통해 성능을 확인하였다. 각 데이터셋의 훈련/테스트셋 및 모델 설정은 [2]와 동일하다.

3.2 실험 결과

제안 방법론의 성능평가를 위해 지도학습과 자기지도학습에서 사용되는 대표적인 모델들과 우수한 성능을 보이는 모델들을 비교모델로 선정하였다. 자기지도학습 모델로는 우선 LR은 가장 전통적인 방법인 Logistic Regression이며, GCN[3], GAT[4], GPRGNN[5]을 선정하였다. 자기지도학습모델로는 MVGRL[6], GRACE[7], Selene[8], HGRL[9], GMI를 선정하였다. 이 중 GPRGNN과 GMI는 각각 지도학습, 자기지도학습에서 가장 우수한 성능을 보이는 모델들이다. 또한 실험에서의 하이퍼파라미터는 δ = 0.9, λ = 1로 설정하였다. 표 2에서 밑줄은 최고 성능을, 볼드체는 자기지도학습 모델내 최고 성능을 의미한다. 실험결과 기존 GMI 대비 모든 데이터셋에서 성능향상을 확인할 수 있었고, Citeseer 데이터셋에서는 지도학습 보다 좋은 결과를 보였다. 이는 GMI와 기존 모델들에 비해 제안 방법론은 단순히 노드 수준에서만 대조적으로 학습하는 것이 아닌, 군집 수준에서의 정보를 학습하므로 군집 사이에 있는 구별이 어려운 노드(마지널 노드)에 대한 표현력이 개선되었기 때문으로 기대할 수 있다.

4. 결 론

본 연구는 그래프 임베딩 시 라벨을 활용하지 않는 자기지도학습 방법을 적용할 때 적용가능한 방법론으로, 기존 네거티브 샘플링을 활용하는 대조학습 방법론들에 군집-노드 간의 소속도 정보인 이웃상호정보량(Neighbor Mutual Information)을 모델 학습 전 간단히 추가함으로써 모델의 표현 성능을 개선할 수 있는 장점이 있다. 제안하는 방법론은 그래프의 전체 군집구조를 활용해 모델이 그래프의 연결구조를 더 잘 학습하게 됨으로써 학습성능을 개선시킬 수 있음을 확인하였다.

References

[1] Liu, Y., Jin, M., Pan, S., Zhou, C., Zheng, Y., Xia, F., & Philip, S. Y., "Graph Self-Supervised Learning: A Survey," *IEEE Transactions on Knowledge and Data Engineering,* Vol. 35, No. 6, pp. 5879-5900, 1 June 2023.

[2] Peng, Z., Huang, W., Luo, M., Zheng, Q., Rong, Y., Xu, T., & Huang, J., Graph representation learning via graphical mutual information maximization. *Proc. of the Web Conference 2020.* pp. 259-270, 2020.

표 2 비교실험 결과
Table 2 Results of experiments

Algorithm		Training data	Dataset			
			Cora	Citeseer	Chameleon	Squirrel
Supervised	LR	$\boldsymbol{X}, \boldsymbol{Y}$	56.6 ± 0.4	57.8 ± 0.2		
	GCN	$\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{Y}$	82.26 ± 1.20	70.19 ± 1.08	54.65 ±2.17	39.50 ± 1.54
	GAT	$\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{Y}$	82.82 ± 0.97	70.62 ± 0.82	55.18 ± 1.95	41.01 ± 0.91
	GPRGNN	$\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{Y}$	85.24 ± 1.14	72.70 ± 0.77	51.93 ± 1.57	35.74 ± 2.35
Unsupervised	MVGRL	$\boldsymbol{X}, \boldsymbol{A}$	84.53 ± 1.05	71.88 ± 0.71	42.34 ± 2.11	33.49 ± 0.84
	GRACE	$\boldsymbol{X}, \boldsymbol{A}$	83.69 ± 0.73	71.37 ± 0.96	45.89 ± 3.10	34.47 ± 1.11
	Selene	$\boldsymbol{X}, \boldsymbol{A}$	56.19 ± 1.52	54.08 ± 1.08	38.93 ± 1.44	OOM
	HGRL	$\boldsymbol{X}, \boldsymbol{A}$	82.08 ± 0.84	70.89 ± 0.75	45.04 ± 1.91	**35.42 ± 0.91**
	GMI	$\boldsymbol{X}, \boldsymbol{A}$	83.80 ± 0.38	74.19 ± 0.44	49.94 ± 1.51	34.51 ± 1.22
	proposed	$\boldsymbol{X}, \boldsymbol{A}$	**84.54** ± 0.36	**75.20** ± 0.42	**52.81** ± 1.76	35.31 ± 0.78

[3] Kipf, T. N., & Welling, M., Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907,* 2016.
[4] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y., Graph attention networks. *arXiv preprint arXiv:1710.10903,* 2017.
[5] Chien, E., Peng, J., Li, P., & Milenkovic, O., Adaptive universal generalized pagerank graph neural network. *arXiv preprint arXiv:2006.07988,* 2020.
[6] Hassani, K., & Khasahmadi, A. H., Contrastive multi-view representation learning on graphs. *International Conference on Machine Learning*. PMLR, pp. 4116-4126. 2020.
[7] Yanqiao, Z., Yichen, X., Feng, Y., Qiang, L., Shu, W., & Liang, W., Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131,* 2020.
[8] Zhong, Z., Gonzalez, G., Grattarola, D., & Pang, J., Unsupervised Heterophilous Network Embedding via r-Ego Network Discrimination. *arXiv preprint arXiv:2203.10866,* 2022.
[9] Chen, J., Zhu, G., Qi, Y., Yuan, C., & Huang, Y., Towards Self-supervised Learning on Graphs with Heterophily. *Proc. of the 31st ACM International Conference on Information & Knowledge Management,* pp. 201-211, 2022.

연 정 흔

2022년 아주대학교 산업공학과 졸업(학사) 2024년 아주대학교 인공지능학과 졸업(석사) 2024년~현재 아주대학교 인공지능학과 박사과정. 관심분야는 자기지도학습, 그래프 기반 알고리즘

신 현 정

2005년 서울대학교 산업공학과 졸업(박사) 2004년~2005년 Max-Planck-Institute for Biological Cybernetics 독일(연구원) 2005년~2006년 Friedrich-MieschерLaboratory, Max-Planck-Institute 독일(수석 연구원) 2006년 서울대학교 의과대학 연구교수 2006년~현재 아주대학교 산업공학과, 데이터사이언스학과, 융합시스템공학과 교수. 관심분야는 머신러닝, 데이터 마이닝, 바이오메디컬 인포매틱스