k-코어 기반 그래프 증강 기법[1][2]

박현석○, 김수지, 박하명†
국민대학교
20151741@kookmin.ac.kr, suji2924@kookmin.ac.kr, hmpark@kookmin.ac.kr

Graph Augmentation Based on k-Core

HyunSeok Park○, Suji Kim, Ha-Myung Park†
Kookmin University

요 약

이 논문에서는 그래프 신경망(Graph Neural Network)의 정확도 향상을 위한 그래프 증강(Graph augmentation)기법을 제안한다. 제안 모델은 k-코어(k-core) 알고리즘(Algorithm)을 통해 가상 정점(Virtual node)을 만드는 모듈(Module)로 구성된다. k-코어 알고리즘으로 중요 노드를 찾고 가상 정점과 연결해주는 과정을 여러 번 거쳐서 새로운 그래프를 만든다. 다양한 공개 데이터셋(Dataset)을 사용한 실험 결과, 제안 모델로 생성한 그래프와 기존 그래프를 함께 학습한 모델이 정점 분류(Node classification)에서 더 높은 정확도를 보임을 확인했다.

1. 서 론

그래프 신경망(Graph Neural Network)의 정확도 향상을 위하여, 어떻게 그래프 증강(Graph augmentation)을 효과적으로 수행할 수 있을까? 그래프 신경망은 그래프의 위상 정보(Graph topology)와 정점 특성(Node feature)을 모두 종합하여 정보를 집계한다. 최근 그래프 신경망 모델은 그래프 구조와 특성(Feature)을 사용해서 추천 시스템[1], 의약 발견[2] 등 다양한 그래프 관련 작업들(Tasks)에서 높은 성능을 보여주고 있다. 데이터 증강(Data augmentation)은 기존의 그래프 데이터를 수정하거나 확장하여 더 유용한 정보를 얻는 방법으로, 인공지능모델(Artificial intelligence model)의 성능 향상을 위해 널리 활용되고 있다. 이미 데이터 증강은 컴퓨터 비전(Computer vision) 및 자연어 처리(Natural language processing) 분야에서 그 중요성을 입증하였으며, 최근 그래프 데이터 분야에도 이를 적용하기 위한 연구들이 진행되고 있으나, 무작위로 증강하는 등 기존 그래프의 연결성을 고려하지 않는 한계가 있다.

본 연구에서는 k-코어 알고리즘[3]을 활용한 그래프 증강 방법을 제안한다. 이는 그래프 내에서 가장 높은 k로 형성된 k-코어를 찾은 후 가상 정점을 생성하고 연결하는 정점 삽입 및 간선 추가 방법이다.

논문의 구성은 다음과 같다. 2장에서는 배경이 된 관련 연구를 소개하고, 3장에서는 그래프를 증강하는 k-코어 모듈(k-core module)을 소개한다, 4장에서 실험에 대해 소개한 후 5장에서 연구를 결론짓는다.

2. 관련 연구

그래프 증강(Graph augmentation): 그래프 증강은 한정된 데이터를 보완하여 모델의 정확도를 올리기 위해서 사용하는 기법이다. 존재하는 그래프 증강 기법으로는 정점 삽입(Node insertion)[4], 간선 교란(Edge perturbation)[5], 그래프 샘플링(Graph sampling)[6]등의 기법들이 존재한다.

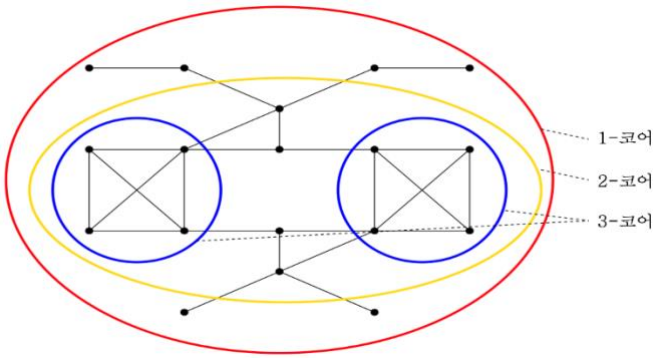

그림 1 k-코어 알고리즘 예시

k-코어(k-core): k-코어 알고리즘(Algorithm)은 그래프에서 정점의 차수(Degree)가 최소 k인 노드들을 찾는 알고리즘이다. 이 알고리즘은 차수가 k보다 작은 노드를 반복적으로 제거하며 진행한다. 이 알고리즘을 통해 그래프 내부에 밀집된 구조를 형성하는 노드를

[1] 이 연구는 기상청 국립기상과학원 「AI 예보지원 및 활용기술 개발」(KMA2021-00123) 의 지원으로 수행되었습니다.
[2] 본 연구는 2022년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음. (2022-0-00964)

찾을 수 있고 커뮤니티 탐지(Community detection), 핵심 정점 탐지(Core node detection)와 같은 문제에 사용된다.

그림 1은 k-코어 알고리즘의 간단한 예시를 보여준다. k=1인 정점은 빨간색 원에만 포함된 정점이고, 파란색 원은 주어진 그래프에서 k=3인 정점을 나타내고, 노란색 원은 k=2인 정점을 나타내고 빨간색 원은 k=1인 정점을 나타낸다.

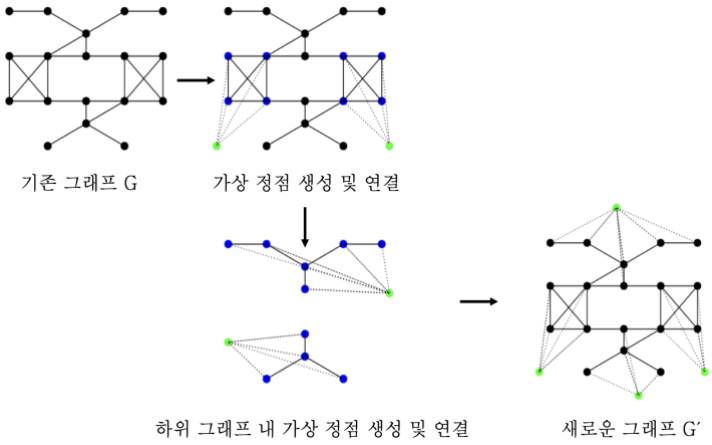

그림 2 증강 그래프 생성방법

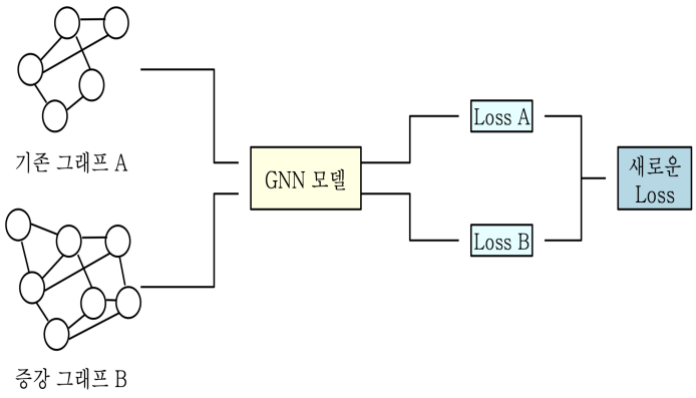

그림 3 제안 모델 구조

3. 제안 기법

그림 2는 제안하는 그래프 증강 방법으로, 그래프 증강을 위해 k-코어 알고리즘을 사용한다. k-코어 알고리즘으로 가장 높은 k를 가진 정점을 찾고, k-코어 정점으로 유도된 서브그래프에서 유니온 파인드(Union-Find) 알고리즘을 사용하여 연결요소(Connected components)를 구한 후, 각각의 연결요소마다 하나의 가상 정점을 생성 후 연결요소 내 모든 정점을 가상 정점과 간선으로 연결한다. 그런 다음, 앞서 계산한 가상 정점과 연결된 정점들을 제외한 하위 그래프에서 같은 기법을 k가 0이 될 때 까지 반복해서 사용한다. 새롭게 생성된 가상 정점의 특성은 가상 정점과 연결된 정점들의 특성의 평균으로 사용한다.

그래프 증강을 통해 새롭게 생성한 그래프는 기존의 그래프와 함께 그래프 신경망 학습에 사용한다. 각각의 학습 결과는 가중 평균하여 손실 함수(Loss function)의 값으로 사용하여 학습을 진행한다. 이 과정에서 가중 평균의 비율 알파(α)를 학습 매개변수(Parameter)로 설정하여 모델과 같이 학습한다. 그림 3은 제안하는 모델의 전반적인 구성을 보여준다.

4. 실 험

4.1 데이터셋 및 모델

제안하는 기법의 정점 분류 작업(Node classification task) 정확도를 확인하기 위해서 논문 인용관계 공개 데이터셋(Dataset)인 Cora, Citeseer, Pubmed[7]를 사용한다. 이 데이터셋들의 정점은 논문을 나타내고 간선은 인용관계, 특성은 논문의 요약에 있는 단어들로 구성되어 있다. 각 데이터셋들의 학습 데이터(Train data)는 전체 노드의 60%, 검증 데이터(Validation data)와 평가 데이터(Test data)는 각각 20%씩 사용한다.

또한 대표적인 그래프 학습 모델인 GCN(Graph Convolution Network)과 GraphSAGE(Graph SAmple and aggreGatE)를 학습에 사용하여 정확도를 평가한다.

4.2 정점 분류(Node classification) 실험

제안하는 기법의 효용성을 입증하기 위해 우리는 두 가지 방법과 제안 기법을 비교한다.

- 기본 방법: 기존 그래프만 사용하여 학습
- 무작위 증강: 가상 정점을 100개 생성하고 기존 그래프에 무작위로 연결하여 증강한 그래프와 기존 그래프를 함께 학습
- k-코어 증강: 제안 기법

표 1 최적 매개변수를 찾기 위한 검색범위

매개변수	검색 범위(Search range)
알파의 학습률 (Learning rate for α)	{0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05}

같은 조건에서 그래프 증강의 정확도 변화를 확인하기 위해 GCN과 GraphSAGE의 학습률은 0.01로 고정하고, 가중 평균 비율 알파(α)의 학습률은 표 1에 명시한 범위로 실험한다. 또한 손실 함수(Loss function)는 크로스 엔트로피(Cross-Entropy) 함수를 사용한다.

표 2 GCN 모델의 정점 분류 정확도

	Cora	Citeseer	Pubmed
기본 방법	85.43	75.44	86.34
무작위 증강	85.31	74.92	85.74
k-코어 증강 (제안 기법)	**85.97**	**75.73**	**86.41**

표 3 GraphSAGE 모델의 정점 분류 정확도

	Cora	Citeseer	Pubmed
기본 방법	87.87	78.15	88.99
무작위 증강	88.22	77.65	88.71
k-코어 증강 (제안 기법)	**88.32**	**78.39**	**89.09**

표 2 와 표 3은 각각 GCN과 GraphSAGE를 학습 모델로 설정하여 정점 분류 작업 정확도를 측정한 결과이다. 결과에 따르면 제안 기법은 기존 그래프만 사용하거나 무작위로 가상 정점을 생성하였을 때보다 모든 데이터셋에서 정확도가 향상하는 것을 확인할 수 있다. 추가로 무작위로 가상 정점을 생성하는 기법의 경우 기존 그래프만 사용했을 때보다 대체로 정확도가 하락하는 양상을 확인할 수 있다. 이 실험 결과를 통해, 제안 기법(k-코어 모듈)을 사용하는 것이 기존 그래프에 적절한 가상 정점을 추가해줌으로써 모델의 정확도를 향상시킨다는 것을 입증한다.

5. 결 론

본 논문에서는 k-코어 모듈을 사용한 그래프 증강 기법을 통해 증강된 그래프와 기존 그래프를 함께 학습하여 가중 평균하는 기법을 제안한다. 제안 기법의 정확도를 평가하기 위하여 세가지 데이터셋에서 GCN 모델과 GraphSAGE 모델을 정점 분류 문제에 적용하여 정확도를 각각 확인한다. 실험 결과, 제안 기법을 통하여 학습을 진행했을 때 정확도가 일관되게 향상되는 것을 확인하였다. 이는 k-코어 모듈을 사용한 그래프 증강 방법이 그래프 학습을 원활하게 할 수 있다는 것을 보여준다.

참고문헌

[1] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec, “Graph convolutional neural networks for web-scale recommender systems.” In Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, 974-983, 2018.

[2] Hanjun Dai, Chengtao Li, Connor Coley, Bo Dai, and Le Song, “Retrosynthesis prediction with conditional graph logic network.” Advances in Neural Information Processing Systems 32, 2019.

[3] Stephen B Seidman. “Network structure and minimum degree.” Social networks 5, 3, 269-287, 1983.

[4] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George EDahl. “Neural message passing for quantum chemistry.” In International conference on machine learning. PMLR, 1263-1272, 2017.

[5] Yuning You, Tianlong Chen, Yang Shen, and Zhangyang Wang, “Graph contrastive learning automated.” In International Conference on Machine Learning. PMLR, 12121-12132, 2021.

[6] Yizhu Jiao, Yun Xiong, Jiawei Zhang, Yao Zhang, Tianqi Zhang, and Yangyong Zhu, “Sub-graph contrast for scalable self-supervised graph representation learning.” In 2020 IEEE international conference on data mining (ICDM). IEEE, 222-231, 2020.

[7] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad, “Collective classification in network data.” AI magazine 29,3, 93-93, 2008.