프로토타입 유사도 기반 그래프 네트워크 설명
Graph Neural Network Explanation Based on Prototype Similarities

Eun-Bi Yoon, Sun-Woo Kim, Yong-Min Shin, Won-Yong Shin
Yonsei University
{yon6286, kswoo9977, jordan3414, wy.shin}@yonsei.ac.kr

요약

그래프 신경망(GNN: graph neural network)은 그래프 마이닝 분야에서 큰 주목을 받고 있지만 내부 결정 과정을 알 수 없는 블랙박스로 간주되었다. 최근 들어 모델의 결정을 설명하는 eXplainable AI(XAI)를 그래프 신경망에 적용하려는 연구가 시도되었다. 본 논문에서는 프로토타입 노드를 도입하여 모델의 결정에 대한 설명을 프로토타입과의 유사도를 기반으로 설명하는 방법론을 제안한다. Infection dataset을 사용하여 제안된 방법론이 모델 결정에 직관적인 설명을 제공함을 실험적으로 보인다.

Ⅰ. 서론

그래프 신경망(GNN: graph neural network)은 그래프 마이닝 분야에서 상당한 성과를 거두었다. 일반적인 신경망 모델은 결정 과정을 해석하기 어려운 블랙박스 모델로 간주한다. 이에 신경망 모델을 해석하기 위한 eXplainable AI(XAI)가 제안이 되었으나 그래프 신경망에 적용하는 연구는 비교적 최근에 시도되었다 [1]. 그래프는 직관적으로 해석 가능하기가 어렵기에 모델 결정에 대한 근거로 이미지 분석에 사용된 heatmap을 제시하는 것은 한계가 있다. 본 논문에서는 그래프 신경망 사용 시 프로토타입 노드를 소개하고 이 노드들 간 유사성 기반 모델의 결정에 대한 설명을 제안한다.

Ⅱ. 방법론

그래프 신경망은 입력 특징 공간 I 내에 노드 v의 입력 특징 I_v와 그래프 (V, E) 의 위상 정보를 바탕으로 latent space에서 최적화된 임베딩을 수행하는데, 이 때 explanation은 latent space에서 제공한다. "Decision cluster D"는 latent space에서 노드의 class를 기준으로 형성된 군집을 의미한다. "Concept C_D"는 각 decision cluster D 내에 모델이 포착한 세부 군집을 의미한다. Decision cluster D 내에 concept C_D의 확률 분포를 가우시안 혼합 모델을 통해 최적화하고, 피팅이 된 각 C_D 분포의 평균을 latent space에서 표현된 C_D의 "프로토타입 노드"로 간주한다. 노드 $v \in D$와 concept C_D 프로토타입과의 유사성은 C_D에 대한 posterior probability $p(C_D|v)$ 로 정의하고, 유사성이 높은 C_D에 대해 $v \in C_D$로 표기한다. C_D의 프로토타입 노드의 입력 특징 F를 weighted average인 $\frac{\sum_{v \in C_D} I_v \cdot p(C_D|v)}{\sum_{v \in C_D} p(C_D|v)}$로 설정하고, 프로토타입의 이웃 노드들은 C_D 내에서 유사성이 높은 노드들의 집합 N_{D,C_D}로 지정한다. 해당 방법론의 explanation은 cluster에 대한 E_{cluster}와 node에 대한 E_{node}로 정의되며 수학적으로 기술하면 다음과 같다.

$$E_{\text{cluster}} : C_D \mapsto \left(F, N_{D,C_D}\right) \text{ s.t } F \in I \text{ and } N_{D,C_D} \subset V, \forall\, C_D$$
$$E_{\text{node}} : (v, D) \mapsto \left(E_{cluster}(C_D), p(C_D|v)\right) \text{ for node } v \in C_D$$

Ⅲ. 실험 설정 및 결과

제안된 방법론의 유효성을 검증하기 위해, synthetic infection dataset [1]을 활용한다. 해당 dataset은 사후 감염여부를 기준으로 2개의 cluster로 분류되며 면역이 없는 노드가 감염자와 접촉된 경우에 사후 감염자로 간주한다. 163개의 노드들과 입력 특징은 총 5개로 초기감염 여부, 면역 여부, 초기비감염 여부, 그리고 두 개의 잡음으로 구성된다. 초기감염자의 노드 인덱스는 0에서 8까지의 9개의 노드들이고 면역자의 노드 인덱스는 9에서 22까지 14개의 노드들이다. 특징 여부는 1로 표기되며 잡음에 대해서는 (0,1) 사이의 임의의 값을 배정한다. 전체 노드의 5%를 학습 데이터로 설정하였고 비감염자 군집(Cluster 0)의 concept 개수는 2, 감염자 군집(Cluster 1)의 concept 개수는 3으로 설정하였다. 실험에는 one-layer GraphSAGE [2]에 tanh를 적용하여 classification layer를 추가한 모델을 사용하였다.

그림 1. Infection dataset에 대한 explanation 결과

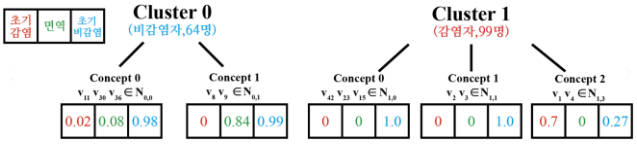

표 1 : 각 프로토타입에 대한 직관적인 해석

프로토타입 해석	Cluster 0	Cluster 1
Concept 0	초기비감염자, 주로 면역자(v_{11})와이웃	초기비감염자, 주로 면역자(v_{15})와 이웃
Concept 1	면역자, 감염자(v_8, v_9)와 이웃	초기비감염자, 감염자(v_2, v_3)와 이웃
Concept 2	없음	초기감염자, 감염자(v_1, v_4)와 이웃

표1은 그림 1에서 얻은 입력 특징 F에 대해 0.5를 기준으로 특징 여부를 판단한 해석을 표기한다. 위 결과를 토대로 E_{node}는 노드 v가 Cluster 1에 속하고 Concept 1에 대해 유사성이 높다면, 초기비감염자이자 감염자와 이웃인 프로토타입과 유사했기 때문이라고 설명한다. 따라서, score를 계산하는 방식이 아닌 모델 내의 설명 단위를 정의하여 직관적인 설명을 제시하는 방법이다. 기존 그래프 신경망 XAI와의 정량적 비교를 수행하기 어려우나, 해당 그래프와 프로토타입들과의 평균 유사성이 $\frac{\sum_D \sum_{C_D} \sum_{v \in C_D} p(C_D|v)}{\text{number of nodes}} = 0.998$로 상당히 높게 나왔고, 각 컨셉 별 프로토타입에 대해 직관적인 해석이 가능하다는 것을 근거로 제안된 방법론이 모델 결정에 직관적인 설명을 제공함을 실험적으로 확인하였다.

ACKNOWLEDGMENT

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C3004345), by the Republic of Korea's MSIT (Ministry of Science and ICT), under the High-Potential Individuals Global Training Program (No. 2020-0-01463) supervised by the IITP (Institute of Information and Communications Technology Planning Evaluation), and by the Yonsei University, Republic of Korea Research Fund of 2021 (2021-22-0083).

참고문헌

[1] F. Baldassarre and H. Azizpour, "Explainability techniques for graph convolutional networks," *arXiv preprint arXiv:1905.13686*, 2019.

[2] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," In *Proc. Advances Neural Inf. Processing Syst. (NIPS)*, Long Beach, CA, Dec. 2017, pp. 1024-1034.