희소한 네트워크에서 부호가 있는 그래프 합성곱 네트워크 방법들의 부호 예측 정확도 분석

김민정 [1], 이연창 [2], 김상욱 [3*]
[1] 한양대학교 인공지능학과 석박통합과정
[2] 조지아공과대학교 컴퓨터공학과 박사후연구원
[3] 한양대학교 컴퓨터소프트웨어학과 교수

kmj0792@hanyang.ac.kr, yeonchang@gatech.edu, wook@hanyang.ac.kr

Analysis of Sign Prediction Accuracy with Signed Graph Convolutional Network Methods in Sparse Networks

Min-Jeong Kim[1], Yeon-Chang Lee[2], Sang-Wook Kim[3*]
[1]Department of Artificial Intelligence, Hanyang University
[2]School of Computational Science and Engineering, Georgia Institute of Technology
[3]Department of Computer Science, Hanyang University

요 약

실세계 네트워크 데이터에서 노드들 간의 관계는 종종 친구/적 혹은 지지/반대와 같이 대조적인 부호를 갖는다. 이러한 네트워크를 분석하기 위해, 부호가 있는 네트워크 임베딩 (signed network embedding, 이하 SNE) 문제에 대한 관심이 급증하고 있다. 특히, 최근 들어 그래프 합성곱 네트워크 기술을 기반으로 하는 SNE 방법들에 대한 연구가 활발히 수행되어 오고 있다. 본 논문에서는, 부호가 있는 네트워크의 희소성 정도가 기존 SNE 방법들의 성능에 어떻게 영향을 미치는 지에 대해 분석하고자 한다. 4 개의 실세계 데이터 집합들을 이용한 실험을 통해, 우리는 기존 방법들의 부호 예측 정확도가 희소한 네트워크들에서는 상당히 감소하는 것을 확인하였다.

1. 서론

소셜 미디어에서 사용자들 간의 관계는 종종 친구/적과 같이 대조적인 부호를 가지며, 이러한 관계는 부호가 있는 네트워크 (signed network) 로 표현될 수 있다 [1]. 이에 따라, 부호가 있는 네트워크 내 노드들을 저차원의 임베딩 벡터로 표현하는 것을 목표로 하는 부호가 있는 네트워크 임베딩 (signed network embedding, 이하 SNE) 문제에 대한 관심이 급증하고 있다 [2]. 특히, 최근에는 그래프 합성곱 네트워크 (graph convolutional networks, 이하 GCN) 기반의 SNE 방법들이 제안되어 왔다 [3, 4, 5, 6, 7].

기존 GCN 기반의 SNE 방법들은 그들의 성능을 평가하기 위해, 공개된 데이터 집합들인 Bitcoin-Alpha, Bitcoin-OTC, Slashdot, Epinions 를 사용하였으며, 이들의 희소성 (sparsity) 은 각각 99.90, 99.94, 99.98, 99.98%이다. 그러나, 실세계 네트워크 데이터들은 앞서 언급한 데이터 집합들에 비해 훨씬 더 희소하다고 알려져 있기 때문에 [5], 기존 GCN 기반의 SNE 방법들이 희소한 네트워크가 주어지더라도 우수한 성능을 보이는지에 대해서 논의될 필요가 있다. 따라서, 본 논문에서 우리는 공개 데이터 집합들의 희소성 정도를 조절해가며, 기존 GCN 기반의 SNE 방법들의 성능이 어떻게 변화하는지 분석해보고자 한다.

2. 기존 GCN 기반의 SNE 방법

GCN 기반의 SNE 방법들은 그래프 내 긍정적인 관계들이 갖는 동질성 (homophily) 과 부정적인 관계들이 갖는 이질성 (heterophily) 을 모델링한다. 또한, 그들은 잘 알려진 사회 이론인 균형 이론과 상태 이론을 활용하여, 네트워크 내 고차 부호 관계 (high-order signed relationships) 의 형성을 분석한다 [6].

구체적으로, SGCN [3] 은 첫 번째 GCN 기반의 SNE 방법이며, SNEA [5] 는 SGCN 에 attention 메커니즘을 추가로 적용한 방법이다. 기본적으로, SGCN 과 SNEA

* 교신저자

는 긍정/부정 관계를 구분하기 위해 각 노드를 두 가지 양/음의 임베딩들로 표현하며, 균형 이론을 통해 이러한 임베딩들을 학습한다.

다음으로, SiGAT [4] 와 SDGNN [6] 은 부호가 있는 네트워크의 방향 정보를 추가적으로 고려한다. 그들은 균형 이론과 상태 이론을 기반으로 motif (예: $\triangle_{i,j,k}, i \rightarrow^{+} j, i \rightarrow^{+} k, k \rightarrow^{+} j$) 를 정의하고, 이러한 motif 들을 학습에 활용한다.

마지막으로, SGCL [7] 은 균형 이론을 활용한 그래프 증강을 통해 양질의 노드 임베딩들을 생성한다. 더 나아가, 대조 학습을 통해 벡터 공간 상에서 긍정적인 관계의 노드들 간 거리는 가깝게, 부정적인 관계의 노드들 간 거리는 멀어지게 학습한다.

3. 실험

실험 환경. 우리는 4 가지 부호가 있는 네트워크 데이터 집합들을 사용하여 실험을 수행한다: Bitcoin-Alpha, Bitcoin-OTC, Slashdot, Epinions. <표 1>은 본 논문에서 사용한 데이터 집합들의 통계를 보여준다.

<표 1> 데이터셋에 대한 통계

데이터 집합	노드	간선	희소성 (%)
Bitcoin-Alpha	3,784	14,145	99.90
Bitcoin-OTC	5,901	21,522	99.94
Slashdot	13,182	36,338	99.98
Epinions	25,148	105,061	99.98

경쟁 방법. 우리는 5 가지 GCN 기반의 SNE 방법들의 성능을 분석한다: SGCN [3], SIGAT [4], SNEA [5], SDGNN [6], SGCL [7]. 공정한 비교를 위해, 우리는 모든 방법들의 임베딩 벡터 차원을 64 로 설정하였다.

실험 방법. 우리는 네트워크의 희소성 정도에 따른 부호 예측 정확도 변화를 확인하는 실험을 수행한다. 부호 예측은 각 방법이 간선 부호를 얼마나 정확하게 분류하는 지를 평가하는 것이다. 이를 위해, 각 데이터 집합을 트레이닝 집합과 테스트 집합으로 분리할 때, 우리는 각 데이터 집합의 간선들 중 x(=80, 60, 40, 20)%를 트레이닝 집합으로 사용하고, 나머지를 테스트 집합으로 사용한다. 즉, x 값이 감소할 수록, 트레이닝 집합의 희소성이 증가한다는 것을 나타낸다.

실험 결과. <표 2>는 GCN 기반 SNE 방법들의 부호 예측 실험을 통해 측정한 area under the curve (AUC) 정확도를 보여준다. <표 2>를 통해, 우리는 x 값이 줄어들면서 대부분 SNE 방법들의 정확도가 감소하는 것을 확인할 수 있다. 이러한 결과는 기존 방법들이 실세계와 유사한 정도의 희소한 네트워크들에서는 우수한 성능을 제공하지 못한다는 것을 의미한다. 다시 말해, 기존 방법들이 실세계 응용에서 광범위하게 활용되기 위해서는, 데이터 희소성을 효과적으로 고려할 수 있어야 한다.

4. 결론

본 논문에서, 우리는 네트워크 희소성이 기존 GCN 기반 SNE 방법들의 성능에 미치는 영향을 분석하였다. 4 개의 데이터 집합들을 이용한 실험을 통해, 대부분의 방법들이 희소한 상황에서는 부호 예측 정확도가 상당히 떨어지는 것을 확인하였다. 이러한 결과는 향후 새로운 SNE 방법을 설계하는 데에 있어서 네트워크의 희소성을 고려해야 한다는 것을 보여준다. 이를 위해, 노드들의 부가 정보 (예: 특성) 나 지식 베이스 등을 추가로 활용할 수 있을 것으로 보인다.

<표 2> 기존 방법들의 간선 부호 예측 정확도

	x	SGCN	SiGAT	SNEA	SDGNN	SGCL
Bitcoin-Alpha	**80**	0.689	0.837	0.805	0.838	0.840
	60	0.702	0.825	0.775	0.835	0.832
	40	0.681	0.809	0.771	0.826	0.827
	20	0.677	0.764	0.700	0.778	0.781
Bitcoin-OTC	**80**	0.763	0.875	0.805	0.876	0.882
	60	0.740	0.871	0.796	0.870	0.873
	40	0.732	0.858	0.804	0.854	0.852
	20	0.672	0.819	0.746	0.825	0.798
Slashdot	**80**	0.805	0.894	0.788	0.876	0.858
	60	0.791	0.892	0.791	0.871	0.843
	40	0.773	0.878	0.779	0.870	0.852
	20	0.777	0.860	0.763	0.853	0.837
Epinions	**80**	0.920	0.961	0.850	0.961	0.947
	60	0.919	0.954	0.837	0.950	0.951
	40	0.921	0.948	0.818	0.948	0.937
	20	0.906	0.927	0.794	0.940	0.922

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))

참고문헌

[1] Jiliang Tang et al., A Survey of Signed Network Mining in Social Media, Computing Surveys, 49, 42, 1–37, 2016.

[2] Pinghua Xu et al., Dual-branch Density Ratio Estimation for Signed Network Embedding, WWW, 2022, 1651–1662.

[3] Tyler Derr et al., Signed Graph Convolutional Networks, ICDM, 2018, 929-934.

[4] Junjie Huang et al., Signed Graph Attention Networks, ICANN, 2019, 566–577.

[5] Yu Li et al., Learning Signed Network Embedding via Graph Attention, AAAI, 2020, 4772–4779.

[6] Junjie Huang et al., SDGNN: Learning Node Representation for Signed Directed Networks, AAAI, 2021, 196–203.

[7] Lin Shu et al., SGCL: Contrastive Representation Learning for Signed Graphs, CIKM, 2021, 1671–1680.