하이퍼엣지 예측 작업에서 네거티브 샘플링 기술의 성능 분석

이다은 [1], 유송경 [1], 고윤용 [2], 김상욱 [3*]
[1] 한양대학교 AI 응용학과 석사과정
[2] 중앙대학교 소프트웨어학부 교수
[3] 한양대학교 컴퓨터소프트웨어학과 교수

ddanable05@hanyang.ac.kr, ssong915@hanyang.ac.kr, yyko@cau.ac.kr, wook@hanyang.ac.kr

Performance Evaluation of Negative Sampling Methods in a Hyperedge Prediction Task

Daeun Lee[1], Songkyung Yu[1], Yunyong Ko[2], Sang-Wook Kim[3*]
[1]Dept. of Artificial Intelligence Application, Hanyang University
[2]School of Computer Science and Engineering, Chung-Ang University
[3]Dept. of Computer Science, Hanyang University

요 약

하이퍼그래프(hypergraph)는 실세계의 여러 객체가 함께 형성하는 복잡한 그룹 관계를 하이퍼엣지(hyperedge)로 정보 손실 없이 모델링할 수 있는 새로운 데이터 구조이다. 하이퍼엣지 예측(hyperedge prediction task)이란 하이퍼그래프로 표현된 실세계 네트워크에서 아직 관찰되지 않은 그룹 관계 혹은 미래에 발생할 가능성이 높은 관계를 예측하는 것으로, 단백질 상호작용 분석(PPI), 추천 시스템, 소셜 네트워크 분석 등 다양한 응용 분야에서 활용된다. 그러나, 하이퍼엣지 예측은 심각한 데이터 희소성 문제로 정확한 예측이 어렵다는 근본적인 한계를 지닌다. 이러한 한계를 완화하기 위해 다양한 네거티브 샘플링(negative sampling) 기술이 활용될 수 있는데, 아직까지 각 샘플링 기술이 하이퍼엣지 예측 정확도에 미치는 효과에 대해 충분히 연구되지 않았다. 본 논문에서는 하이퍼엣지 예측에 활용되는 다양한 네거티브 샘플링 방법의 효과를 분석한다. 실험 결과를 통해, 네거티브 샘플링 기법과 포지티브와 네거티브 하이퍼엣지 수의 비율에 따른 정확도 변화 양상을 분석한다.

1. 서론

실세계에는 세 개 이상의 객체들이 함께 형성하는 다양한 복잡한 그룹 관계들이 존재한다 [1, 2, 3]. 예를 들어, 특정 질병을 유발하는 단백질들의 관계, 사용자가 함께 구매한 물품들의 관계, 공동으로 논문을 집필한 연구자들의 관계 등이 있다.

그래프 (graph)는 노드와 엣지로 구성되는 데이터 구조로, 노드는 객체를, 엣지는 두 노드 간의 관계를 나타낸다. 그래프를 이용하여 앞서 기술한 실세계의 그룹 관계를 모델링할 경우, 원본 그룹 정보가 불가피하게 손실될 수밖에 없다는 한계를 가진다 [4, 5, 6].

반면, 하이퍼그래프(hypergraph)는 임의의 수의 노드들의 관계를 하나의 하이퍼엣지(hyperedge)를 통해 모델링하는 일반화된 그래프 데이터 구조로, 정보 손실 없이 실세계의 그룹 관계 정보를 표현할 수 있다 [7]. 이러한 하이퍼그래프의 우수한 표현력 덕분에 하이퍼그래프 기반의 네트워크 학습 기술들이 최근 활발하게 연구되고 있다 [8, 9].

하이퍼그래프 학습 기술의 대표적 다운 스트림 작업(downstream task)으로 하이퍼엣지 예측 (hyperedge prediction) 작업이 있다. 이는 하이퍼그래프에서 아직 관찰되지 않았거나, 또는 미래에 형성될 가능성이 높은 그룹 관계를 예측하는 작업을 의미한다.

이러한 하이퍼엣지 예측 작업은 소셜 네트워크 분석, 추천 시스템, 생명정보학 등 다양한 실세계 응용에 적용이 가능하다 [10, 11, 12, 13]. 예를 들어, 단백질 상호작용 네트워크를 분석하여 잠재적으로 새로운 질병을 유발할 수 있는 단백질 그룹을 예측 및 분석함

* 교신 저자

으로써, 분석 결과를 해당 질병의 예방 및 치료를 위한 신약 개발에 활용될 수 있다.

그러나 하이퍼그래프 기반의 학습 기술은 심각한 데이터 희소성 (data sparsity) 문제를 가지고 있다[14]. 이러한 문제를 완화하기 위해, 실제 존재하지 않는 그룹 관계를 학습 과정에서 추가적으로 사용하고자 하는 네거티브 하이퍼엣지 샘플링 (negative sampling) 기술들이 연구되고 있다 [14, 15, 16, 17].

샘플링된 네거티브 하이퍼엣지는 모델 학습 과정에서 모델이 실제 존재하는 하이퍼엣지 (즉, 포지티브 하이퍼엣지)와 그렇지 않은 하이퍼엣지를 구분하는 능력을 향상시키는데 도움을 준다고 알려져 있다. 특히 '어려운' 네거티브 샘플일수록 모델의 학습에 제공하는 정보의 양이 많은 경향이 있다 [17].

그러나, 다양한 네거티브 하이퍼엣지 샘플링 방법들 중 1) 어떠한 방식이 가장 효과적인지, 2) 학습 시, 가장 효과적인 포지티브 하이퍼엣지 와 네거티브 하이퍼엣지의 비율은 얼마인지에 대한 연구는 충분히 이뤄지지 않았다. 이러한 동기에서 본 논문은 하이퍼엣지 예측 시, 각 네거티브 샘플링 방법에 대한 효과에 대해 검증하고, 더불어, 포지티브 하이퍼엣지와 네거티브 하이퍼엣지의 비율에 따른 하이퍼엣지 예측 정확도를 분석하고자 한다.

2. 네거티브 샘플링

네거티브 샘플링이란 실제 존재하지 않는 가상의 데이터를 생성하는 기술로, 네거티브 샘플링에서 의해 생성된 네거티브 데이터 (예: 네거티브 하이퍼엣지)는 모델의 학습 과정에 음의 정보를 제공한다. 네거티브 샘플링은 그래프 학습, 자연어처리, 컴퓨터 비전 등 많은 영역에서 머신 러닝 모델의 정확도를 개선하는 데 사용된다 [18, 19, 20].

구체적으로, 네거티브 샘플을 활용하여 머신 러닝 모델을 학습할 때, 손실함수는 아래와 같이 정의된다. $x+$는 포지티브 샘플, $x-$는 네거티브 샘플을 의미한다 [21].

$$\text{Loss} = -[\log \sigma(S(f(x), f(x+))) + \log(\sigma(-S(f(x), f(x-)))]$$

하이퍼엣지 예측에서의 네거티브 샘플은 네거티브 하이퍼엣지이다. 네거티브 하이퍼엣지는 실제 관측되지 않은 그룹 관계 정보를 나타낸다. 따라서, 포지티브 하이퍼엣지와 노드 구성의 차이가 존재하고, 이로 인해 하이퍼엣지 특징 정보가 다르게 표현된다.

이 점을 활용하여 하이퍼엣지 예측에서 위 손실함수는 포지티브 하이퍼엣지 특징과 네거티브 하이퍼엣지 특징을 고려하여, 이 둘을 잘 구별하도록 학습하는 것을 목표로 한다.

기존 연구에서는 네거티브 샘플링 방법으로, 휴리스틱 방법을 주로 사용한다. 휴리스틱 방법에는 하이퍼엣지 크기 분포와 포지티브 하이퍼엣지 구성 노드를 유지하는 비율, 대체 노드 선택 방법에 따라 다양한 샘플링 방법들이 존재한다.

본 논문에서는 휴리스틱 방법들 중 최신 방법[14]인 SNS, MNS, CNS 방법을 중점적으로 다룬다. 이 방법들은 모두 포지티브 하이퍼엣지 크기 분포에 따라 네거티브 하이퍼엣지의 크기(k)를 결정하며, 각 방법은 포지티브 하이퍼엣지 구성 노드 유지 비율과 대체 노드 선택 방법에 따라 차이점이 존재한다. 각 방법에 대한 설명은 아래와 같다.

- SNS 는 k 개의 노드를 모두 랜덤하게 선택하여 하이퍼엣지를 구성하는 방법이다.
- MNS 는 클릭 확장(clique expansion)을 통해 하이퍼그래프를 그래프로 표현하고, 그래프 기반 이웃 정보를 기반으로 노드를 선택하는 방법이다. 우선, 포지티브 하이퍼엣지 내 2 개의 노드 (기준 노드)를 선택하고, 그래프로 표현했을 때의 기준 노드 이웃 중 하나의 노드를 선택하는 과정(선택된 노드는 기준 노드에 추가)을 반복한다. 이를 통해 k 개의 노드 집합을 구성한다.
- CNS 는 포지티브 하이퍼엣지 내 1 개의 노드를 랜덤하게 대체하여 k 크기의 하이퍼엣지를 구성하는 방법이다.

이 방법들을 활용한 많은 기존 연구들은 논문 별로, 포지티브 하이퍼엣지와 네거티브 하이퍼엣지 수의 비율을 임의로 정하여 학습을 진행한다.

따라서 기존 연구에서는 논문마다 비율에 따른 실험 환경이 다르며, 포지티브 하이퍼엣지와 네거티브 하이퍼엣지 간 비율이 학습에 미치는 효과에 대해 충분히 연구된 적 없음을 알 수 있다.

본 논문에서는 실험을 통해 하이퍼엣지 예측 시, 네거티브 샘플링에 대한 분석을 진행한다.

- 예측 성능: 우리는 각 네거티브 하이퍼엣지 샘플링의 학습 성능을 비교하며 가장 효과적인 샘플링 방법에 대해 분석한다.
- 샘플링 비율에 따른 성능 변화: 우리는 포지티브 하이퍼엣지 수와 네거티브 하이퍼엣지 수의 비율에 따른 학습 성능을 분석하여 가장 효과적인 비율에 대해 보인다.

3. 실험

3.1 실험 환경

본 논문에서는 co-citation 데이터 셋인 [1]citeseer 를 사용하여 실험을 진행하였다. co-citation 데이터 셋에

서 각 노드는 paper 이며, 하이퍼엣지는 하나의 paper 를 인용한 papers 이다. 표 1 은 citeseer 데이터 셋의 통계를 나타낸다.

<표 1> Citeseer 의 데이터 셋 통계

	Citeseer
Number of nodes	1457
Number of edges	1078
Average size of hyperedges	3.2
Minimum size of hyperedges	2
Maximum size of hyperedges	26
Dimension of node features	3703

하이퍼그래프 학습 모델은 HNHN [6] 을 사용하였다. HNHN 은 스타 확장(star expansion)을 활용하여, 노드와 하이퍼엣지 임베딩을 업데이트함으로써 그룹 관계 정보를 반영하는 모델이다.

또한, 하이퍼엣지 예측은 하이퍼엣지를 구성하는 노드 정보를 기반으로 하이퍼엣지 특징을 정의하여, 하이퍼엣지의 존재 여부를 예측하는 문제이다.

따라서 본 논문에서는 노드 특징을 합쳐 하이퍼엣지 특징을 만드는 방법으로, 최대/최소 방법을 (maxmin aggregator)를 사용하였다 [13]. 이는 하이퍼엣지를 구성하는 노드들의 특징에 대해 (element-wise maximum - element-wise minimum)을 활용하는 방식이다.

3.2 실험 방법

학습을 위해 데이터셋을 train:val:test = 6:2:2 로 분할하여 사용하였으며, 학습과 테스트 시 사용한 네거티브 샘플링은 분할된 학습 데이터셋과 테스트 데이터셋을 기반으로 생성하였다.

하이퍼엣지 예측 시, 각 네거티브 샘플링 방법의 효과를 보기 위한 평가 지표로는 Average Precision (AP)를 사용하였으며, 5 번 실험의 결과 평균을 사용하였다. 실험은 총 3 개의 샘플링 방법 (SNS, MNS, CNS)에 대해 진행하였으며, 학습 시, 포지티브 하이퍼엣지 수와의 비율은 1:0/1/2/3/4 (포지티브: 네거티브)로 총 5 개의 비율에 대해 실험하였다.

3.3 실험 결과

표 2 는 네거티브 샘플링을 사용하지 않았을 때의 결과이고, 표 3, 4, 5 는 학습 시, 사용한 네거티브 샘플링 방법에 따른 결과 및 포지티브 하이퍼엣지 수와 네거티브 하이퍼엣지 수의 비율에 따른 결과를 나타낸 것이다.

표 2 의 실험 결과를 통해, 네거티브 샘플링을 사용하는 것이 학습 성능 개선에 도움을 줌을 확인하였다.

표 3-5 의 실험 결과를 통해, SNS 기반 네거티브 샘플로 테스트한 실험 환경을 제외하면, MNS 기반 네거티브 하이퍼엣지를 사용하여 학습한 모델이 가장 높은 성능을 보임을 확인하였다.

이를 통해, SNS 방법과 같이 노드 구성을 전체로 바꾸거나 CNS 방법과 같이 포지티브 하이퍼엣지 구성 중 하나의 노드만을 대체하는 것이 아닌, 포지티브 하이퍼엣지 기반 이웃 노드들로 네거티브 하이퍼엣지를 구성하는 MNS 방법이 학습에 가장 효과적임을 알 수 있다.

또한, CNS 기반 네거티브 하이퍼엣지로 학습한 모델의 성능이 가장 좋지 않음을 확인하였다.

이는 하이퍼엣지 특징을 만드는 최대/최소 방법에서, CNS 기반으로 만든 네거티브 하이퍼엣지의 경우, 포지티브 하이퍼엣지와 특징 차이가 크지 않을 가능성이 존재한다. 따라서 학습에 어려움이 존재함을 알 수 있다.

더불어, 포지티브 하이퍼엣지와 네거티브 하이퍼엣지의 비율을 1:2 로 학습했을 때, 가장 성능이 좋거나 최고 성능과 유사한 성능을 보임을 확인하였다.

<표 2> 네거티브 샘플링 없이 학습한 결과

	SNS	MNS	CNS	All
1:0	0.4479	0.4514	0.4672	0.4547

<표 3> SNS 기반 네거티브 하이퍼엣지로 학습한 결과

	SNS	MNS	CNS	All
1:1	0.8516	0.7544	0.6027	0.7208
1:2	**0.8847**	**0.7614**	**0.6064**	**0.7293**
1:3	0.7781	0.7004	0.5746	0.6592
1:4	0.7469	0.6662	0.5603	0.6417

<표 4> MNS 기반 네거티브 하이퍼엣지로 학습한 결과

	SNS	MNS	CNS	All
1:1	0.8161	0.7230	0.6048	0.7004
1:2	0.8557	0.7666	0.6046	0.7194
1:3	**0.8704**	**0.7789**	**0.6209**	**0.7376**
1:4	0.8050	0.7061	0.5871	0.6786

<표 5> CNS 기반 네거티브 하이퍼엣지로 학습한 결과

	SNS	MNS	CNS	All
1:1	0.5961	0.5300	0.5043	0.5249
1:2	**0.7425**	**0.6307**	**0.5408**	**0.6189**
1:3	0.7112	0.5849	0.5408	0.6063
1:4	0.5128	0.5276	0.5030	0.5018

4. 결론 및 향후 연구

본 논문은 하이퍼엣지 예측 정확도 개선을 위해 네

[1] https://linqs.soe.ucsc.edu/data

거티브 하이퍼엣지 샘플링 방법별 효과를 확인하고, 네거티브 하이퍼엣지 활용 시, 가장 효과적인 포지티브 하이퍼엣지 수와 네거티브 하이퍼엣지 수의 비율을 분석하였다.

실험 결과를 통해 다양한 네거티브 하이퍼엣지 샘플링 방법 중 MNS 가 가장 효과적임을 확인할 수 있다.

또한, 네거티브 샘플링을 활용하지 않는 경우와의 비교를 통해 네거티브 샘플링을 활용하는 것 자체만으로도 학습 정확도를 높일 수 있다는 결론을 얻을 수 있다.

더불어, 포지티브 하이퍼엣지 수 대비 네거티브 하이퍼엣지 수를 2 배로 하였을 때 가장 효과적인 학습이 될 수 있음을 확인하였다.

이러한 결과를 바탕으로, MNS 방법을 통해 생성되는 네거티브 하이퍼엣지에 대한 추가적인 분석을 진행하여, 휴리스틱 방법이 아닌, 고도화된 네거티브 하이퍼엣지 샘플링 방법을 개발하는 것이 향후 연구 과제이다.

사사

이 논문은 (1)정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No. 2020-0-01373, 인공지능대학원지원(한양대학교))과 (2)한국연구재단의 지원(No.2018R1A5A7059549)을 받아 수행된 연구임. 또한, (3) 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2022-00155586, 실세계의 다양한 다운스트림 태스크를 위한 고성능 빅 하이퍼그래프 마이닝 플랫폼 개발(SW 스타랩)

참고문헌

[1] Austin R Benson, Rediet Abebe, Michael T Schaub, Ali Jadbabaie, and Jon Kleinberg. 2018. Simplicial closure and higher-order link prediction. PNAS 115, 48 (2018), E11221–E11230.

[2] Leo Torres, Ann S Blevins, Danielle Bassett, and Tina Eliassi-Rad. 2021. The why, how, and when of representations for complex systems. SIAM Rev. 63, 3 (2021), 435–485

[3] S. Klamt, U.-U. Haus, and F. Theis, “Hypergraphs and cellular networks,” PLoS Comput. Biol., vol. 5, no. 5, May 2009 Art. no. e1000385.

[4] Grover, Aditya, and Jure Leskovec. "node2vec: Scalable feature learning for networks." Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016.

[5] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. 2019. Hypergraph neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence. 3558–3565

[6] Yihe Dong, Will Sawin, and Yoshua Bengio. 2020. HNHN: hypergraph networks with hyperedge neurons. arXiv preprint arXiv:2006.12278 (2020)

[7] M. M. Wolf, A. M. Klinvex, and D. M. Dunlavy, “Advantages to modeling relational data using hypergraphs versus graphs,” in Proc.IEEE High Perform. Extreme Comput. Conf. (HPEC), Sep. 2016, pp. 1–7.

[8] Fan, Wenqi, et al. "Graph neural networks for social recommendation." The world wide web conference. 2019.

[9] Fan, Wenfei. "Graph pattern matching revised for social network analysis." Proceedings of the 15th international conference on database theory. 2012.

[10] Zheng Liu, Xing Xie, and Lei Chen. 2018. Context-aware academic collaborator recommendation. In KDD

[11] Chia-An Yu, Ching-Lun Tai, Tak-Shing Chan, and Yi-Hsuan Yang. 2018. Modeling multi-way relations with hypergraph embedding. In CIKM

[12] Muhan Zhang, Zhicheng Cui, Shali Jiang, and Yixin Chen. 2018. Beyond link prediction: Predicting hyperlinks in adjacency space. In AAAI.

[13] Naganand Yadati, Vikram Nitin, Madhav Nimishakavi, Prateek Yadav, Anand Louis, and Partha Talukdar. 2020. NHP: Neural Hypergraph Link Prediction. In CIKM

[14] Prasanna Patil, Govind Sharma, and M Narasimha Murty. 2020. Negative sampling for hyperlink prediction in networks. In PAKDD

[15] Se-eun Yoon, Hyungseok Song, Kijung Shin, and Yung Yi. 2020. How Much and When Do We Need Higher-order Information in Hypergraphs? A Case Study on Hyperedge Prediction. In WWW

[16] Ruochi Zhang, Yuesong Zou, and Jian Ma. 2020. Hyper-SAGNN: a self-attention based graph neural network for hypergraphs. In ICLR

[17] Hyunjin Hwang, Seungwoo Lee, Chanyoung Park, and Kijung Shin. 2022. AHP:Learning to Negative Sample for Hyperedge Prediction. In Proc. of the ACM International Conference on Research & Development in Information Retrieval (ACM SIGIR). 2237–2242

[18] Z. Yang, M. Ding, C. Zhou, H. Yang, J. Zhou, and J. Tang, “Understanding negative sampling in graph representation learning,” in KDD, 2020, pp. 1666–1676.

[19] X. Wu, C. Gao, L. Zang, J. Han, Z. Wang, and S. Hu, “Esimcse: Enhanced sample building method for contrastive learning of unsupervised sentence embedding,” arXiv preprint arXiv:2109.04380, 2021.

[20] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, and D. Larlus, “Hard negative mixing for contrastive learning,” NIPS, vol. 33, pp. 21 798–21 809, 2020.

[21] Z. Yang et al., "Does Negative Sampling Matter? A Review with Insights into its Theory and Applications," in IEEE Transactions on Pattern Analysis and Machine Intelligence