Node2vec을 이용한 drug repositioning

이신의*, 하지환*, 박상현*

*연세대학교 컴퓨터과학과

e-mail : {lsnfamily02, jihwanha, sanghyun }@yonsei.ac.kr

Drug repositioning with node2vec

Seanie Lee*, Jihwan Ha*, Sanghyun Park*

*Dept. of Computer Science, Yonsei University

요 약

최근 신약개발의 새로운 방법론으로 신약재창출(drug repositioning)이 각광 받고 있다. 이 연구에서는 그래프 엠베딩(graph embedding)을 이용하여 약물의 새로운 치료효과를 발견하고자 한다. 약물-질병 이분 네트워크를 구축하고, 이를 node2vec 알고리즘을 이용하여 그래프의 정점을 벡터로 표현한다. 이를 바탕으로, 약물 정점과 질병 정점의 유사도를 계산하여 두 정점 사이의 관계를 추론함으로써 새로운 약물-질병 관계를 발견한다. 실험 결과 높은 정확률을 보이는 것으로 나타났다

1. 서 론

신약 개발은 많은 비용과 시간이 소요된다. 이전 연구 보고에 따르면, 신약 개발에 평균적으로 15년이 걸리며 8억달러에서 15달러의 비용이 투입된다[1,2]. 그에 비해 정식으로 승인 받은 약물의 수는 현저히 적다. 이를 타개하고자 이전에 개발된 약을 다른 치료 목적으로 사용하는 신약재창출(drug repositioning)이 시도되고 있으며, 상당한 성과를 보이고 있다.

약물-질병 네트워크를 기반으로 새로운 약물-질병 관계를 추론하는 방법론들이 활발히 연구되고 있다. [3]은 약물-질병 이분네트워크(bipartite network)를 구축하였다. 약물이 질병을 치료할 수 있는지를 점수로 산출하여, 약물이 치료할 수 있는 새로운 질병들을 추천하는 알고리즘을 제안하였다.

[4]는 질병 네트워크와 약물 네트워크를 연결하는 이종 네트워크(heterogeneous network)를 구축하였다. 질병 간의 유사도를 계산하여, 두 질병 정점의 간선 가중치로 사용하였다. 이와 유사하게 약물 유사도를 구하여 약물 정점 사이의 간선 가중치로 이용하였다. 최종적으로 약물이 특정 질병을 치료하는 효과가 있다면, 두 질병과 약물이 간선으로 연결되어, 서로 다른 두 네트워크가 연결된다. 거대한 네트워크에서 random walk를 하며 새로운 질병-약물 관계를 예측하였다.

[5]에서는 약물-질병 인과 네트워크를 구축하여 약물-질병 관계 점수를 측정하였다. 그러나 이러한 인과관계에서는 약물이 질병을 치료하는 것인지, 악영향을 미치는지 파악할 수 없다는 단점이 있다. 이를 보완하고자, PMF(Probabilistic Matrix Factorization)에서 치료효과 관계인 약물-질병 관계만을 추출하였다.

이 연구에서는 이전과는 다른 접근방법을 제시하고 제시하고자 한다. node2vec[6]을 이용하여 새로운 약물-질병 관계를 예측하였다. 약물-질병 네트워크를 구축하고, 각 정점을 D 차원의 벡터로 표현한다. 약물 벡터와 질병 벡터의 코사인 유사도를 계산하여 일정 임계치보다 높은 관계를 치료관계로 간주하였다. 논문의 구성은 다음과 같다. 2.1 에서는 데이터의 수집과 통계를 기술하고, 2.2 에서는 실험 환경과 구체적인 결과를 논의하고자 한다. 마지막 3.1 에서는 후속 연구로 성능 개선 방향에 대해 서술한다.

2. 데이터 출처 및 실험 방법

2.1 데이터 출처

DrugBank 에서 제공하는 모든 약물(10500 개)과 약물이 치료로 하는 질병 쌍 크롤링하였다. DrugBank 데이터베이스에서 삭제된 약물 7 개와 질병 정보가 누락된 약물 8830 개는 데이터에서 제외하였다. 최종적으로 1662 개의 약물과 2852 개의 질병, 그리고 6732 개의 약물-질병 쌍이 실험 데이터로 사용되었다.

<표 1> DrugBank 데이터 통계

총 질병 수	총 약물 수	질병-약물 쌍
2852	1662	6732

2.2 실험 방법 및 결과

DrugBank 에서 수집한 각각의 약물과 질병을 네트워크의 정점으로 하며, 한 약물이 특정 질병에 치료목적으로 사용된다면 두 정점 사이를 간선으로 잇는다. 최종적으로 약물-질병 네트워크가 구축된다. (그림 1)

학습 데이터와 테스트 데이터는 다음과 같이 구성하였다.

알려진 약물-질병 간선의 80%를 훈련 데이터로 하고, 나머지 20%의 간선은 제거하고 그래프를 구축하였다. 제거된 간선과 더불어 서로 연결되지 않은 약물-질병 쌍을 같은 수만 큼 구성하여 테스트 데이터로 사용하였다.

학습 데이터로 구축한 이분 그래프 G =(V, E)에 대하여 node2vec 알고리즘을 이용하여 모든 노드를 벡터로 표현한다. Node2vec 은 BFS(Breadth First Search)와 DFS(Depth First Search)를 적절히 섞어가며 그래프의 정점을 샘플링 한다. 두 탐색 방법을 어떻게 조합할지는 매개변수 p 와 q 로 결정하게 된다. 그래프의 연결성을 벡터 정보에 넣어야 하기 때문에, DFS 보다 BFS 를 비중 있게 두었다. 샘플링 한 노드 중 에서 같은 커뮤니티에 속하는 노드 또는 네트워크 상에서 같은 역할을 하는 노드끼리 벡터 공간 상에서 가까이 위치하도록 학습한다. 최종적으로 약물 벡터와 질병 벡터의 코사인 유사도를 계산하여 일정 임계치보다 높은 쌍은 약물이 질병을 치료한다고 예측하였다. 현재까지 알려진 약물 질병 관계만을 기반으로 정답을 구성하였다. 최종적으로 정확률, 재현율, 그리고 F1 score 를 구하였다. (표 2)

실험결과 높은 정확률을 보이는 반면, 상당히 낮은 재현율을 보였다. 질병 정점과 약물 정점 사이에 간선이 없는 관계는 정확히 예측한 반면, 두 정점 사이에 간선이 존재하는 경우 예측이 부정확한 경우가 상당히 많았다. 이를 개선하기 위해서는 최적의 모델 매개변수와 유사도 임계치를 찾아야 할 것이다.

<표 2. 실험 결과>

정확률	재현율	F1 score
0.90	0.42	0.57

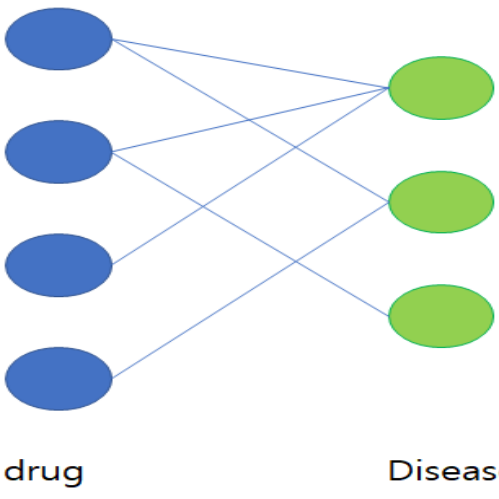

그림 1. 약물-질병 이분 그래프. 가중치가 없는 이분 그래프

결론 및 향후 연구

실험 결과 높은 정확률을 보였으나 재현율이 현저하게 떨어짐을 확인할 수 있었다. 향후연구에는 정확률을 유지하면서 재현율을 높이는 방안을 고려해야 할 것이다. 첫째로 모델의 파라미터 최적화이다. Node2vec 의 여러 파라미터 값에 따라 그래프의 임베딩 값이 달라지게 되며, 이것이 간선 예측에 큰 영향을 미친다. 네트워크 상에서 두 정점의 연결성을 가장 잘 반영할 수 있는 정점의 벡터값을 학습해야 할 것이다. 현재는 임의로 매개변수를 설정하였다. 그러나 좀더 체계적으로 이를 탐색해야 할 것이다. 가능한 방법으로는 grid search 가 있을 것이다. 가능한 매개변수의 후보군을 설정하고, validation set 에서 가장 성능이 높은 매개변수를 찾는 것이다. 또한 정확률 재현율이 유사도의 임계치 값에 크게 많이 달라진다. 이 또한 마찬가지로 K-fold cross validation 을 이용하여 최적의 값을 찾아야 할 것이다.

두번째로 머신러닝 기법을 이용하여 약물과 질병의 관계를 예측하는 것이다. 현재는 유사도 임계치를 사용하여 두 정점 사이의 간선 존재 유무를 확인 하였다. 이를 머신러닝을 이용하여 자동화 할 수 있을 것이다. Node2vec 을 이용하여 질병-약물 네트워크 상의 노드 특징(feature)를 학습한다. 그리고 이 특징(feature)를 input feature 로 하여, SVM 이나 Neural Network classifier 를 훈련시킨다. 최종적으로 테스트 데이터에서 두 질병과 약물 사이의 관계를 예측한다.

마지막으로 약물-질병 관계 정보뿐만 아니라 질병 유사도, 약물의 화학적 유사도 정보를 이용하는 것이다. 선행연구에서 언급하였듯이, 질병 유사도를 이용하여 질병 네트워크를, 약물 유사도를 이용하여 약물 네트워크를 구축한다. 약물 유사도는 DrugBank 에서 제공하는 약물 간의 화학적 유사도를 이용한다. 질병 유사도는 MimMiner[7]에서 제공하는 두 질병 간의 유사정보를 활용할 것이다. 또 다른 방법으로는 단어 word2vec[8]을 이용하는 것이다. 질병 이름이 언급되는 모든 논문의 초록을 PubMed 에서 수집한다. 그리고 질병 이름을 단어 벡터로 학습시키고, 두 단어 벡터의 코사인 유사도를 두 질병의 유사도로 이용한다. 두 네트워크를 약물-질병 관계 정보를 이용하여 연결한다. 추가적인 정보를 활용하여 더 복잡한 네트워크를 구축한다면, 약물-질병 관계 예측의 정확도를 향상시킬 수 있을 것이다.

참고문헌

[1] DiMasi, Joseph A. "New drug development in the United States from 1963 to 1999." *Clinical Pharmacology & Therapeutics* 69.5 (2001): 286-296.

[2] Adams, Christopher P., and Van V. Brantner. "Estimating the cost of new drug development: is it really $802 million?." *Health affairs* 25.2 (2006): 420-428.

[3] Chen, Hailin, et al. "Network-based inference methods for drug repositioning." *Computational and mathematical methods in medicine* 2015 (2015).

[4] Luo, Huimin, et al. "Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm." *Bioinformatics* 32.17 (2016): 2664-2671.

[5] Yang, Jihong, et al. "Drug-disease association and drug-repositioning predictions in complex diseases using causal inference-probabilistic matrix factorization." *Journal of chemical information and modeling* 54.9 (2014): 2562-2569.

[6] 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*. ACM, 2016.

[7] Van Driel, Marc A., et al. "A text-mining analysis of the human phenome." *European journal of human genetics* 14.5 (2006): 535.

[8] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*. 2013.