암 유전체 데이터를 효과적으로 학습하기 위한 Node2Vec 기반의 새로운 2 차원 이미지 표현기법

최종환, 박상현†
연세대학교 컴퓨터과학과
e-mail: {mathcombio, sanghyun}@yonsei.ac.kr

A novel Node2Vec-based 2-D image representation method for effective learning of cancer genomic data

Jonghwan Choi, Sanghyun Park†
Dept. of Computer Science, Yonsei University

요 약

4 차산업혁명의 발달은 전 세계가 건강한 삶에 관련된 스마트시티 및 맞춤형 치료에 큰 관심을 갖게 하였고, 특히 기계학습 기술은 암을 극복하기 위한 유전체 기반의 정밀 의학 연구에 널리 활용되고 있어 암환자의 예후 예측 및 예후에 따른 맞춤형 치료 전략 수립 등을 가능케하였다. 하지만 암 예후 예측 연구에 주로 사용되는 유전자 발현량 데이터는 약 17,000 개의 유전자를 갖는 반면에 샘플의 수가 200 여개 밖에 없는 문제를 안고 있어, 예후 예측을 위한 신경망 모델의 일반화를 어렵게 한다. 이러한 문제를 해결하기 위해 본 연구에서는 고차원의 유전자 발현량 데이터를 신경망 모델이 효과적으로 학습할 수 있도록 2D 이미지로 표현하는 기법을 제안한다. 길이 17,000 인 1 차원 유전자 벡터를 64x64 크기의 2 차원 이미지로 사상하여 입력크기를 압축하였다. 2 차원 평면 상의 유전자 좌표를 구하기 위해 유전자 네트워크 데이터와 Node2Vec 이 활용되었고, 이미지 기반의 암 예후 예측을 수행하기 위해 합성곱 신경망 모델을 사용하였다. 제안하는 기법을 정확하게 평가하기 위해 이중 교차 검증 및 무작위 탐색 기법으로 모델 선택 및 평가 작업을 수행하였고, 그 결과로 베이스라인 모델인 고차원의 유전자 벡터를 입력 받는 다층 퍼셉트론 모델보다 더 높은 예측 정확도를 보여주는 것을 확인하였다.

1. 서론

기계학습을 포함한 4 차산업혁명은 스마트시티(smart city), 맞춤형 의학(personalized medicine) 등 사람의 건강에 관련된 기술의 발달을 일으켰고, 특히 생물정보학(bioinformatics) 분야에 큰 패러다임을 일으켜 발암 기전(oncogenesis)을 연구하기 위한 기계학습(machine learning) 기반의 암 진단(diagnosis) 또는 예후(prognosis) 예측 모델의 연구가 활발해지도록 촉진하였다[1,2]. 하지만, 2018 년 12 월에 발표된 보건복지부의 국가암등록통계 발표자료[3]에 의하면, 국내 폐암, 간암 그리고 췌장암 환자의 5 년 생존율은 각각 28.2%, 34.6%, 11.4%로 매우 나쁜 예후를 보이고 있어, 암환자의 생존 및 치료를 위해 보다 정확한 예후 예측 모델의 연구 및 개발이 필요한 실정이다.

암 예후 예측 연구에 많이 사용되는 자료로 암환자의 유전자 발현량 데이터(gene expression profiling)가 있다[4]. 유전자 발현량 데이터는 사람이 지닌 2 만여 개 암호화 유전자(protein-coding gene)의 활성화 정도를 측정한 수치형 자료(numeric data)로, 이를 분석하기 위해 다양한 기계학습 기반의 예후 예측 모델들이 제안되었다[2]. 하지만 유전체 데이터는 특징의 개수가 몹시 많은 반면에 샘플의 수가 극히 적은 차원의 저주(curse of dimensionality) 문제를 가지고 있어, 기계학습 및 신경망 모델의 일반화 성능 저하를 방지하기 위한 차원 축소 기법이 필요하다[5].

유전자 발현량 데이터에 효과적인 차원 축소 전략을 세우기 위해 단백질-단백질 상호작용(protein-protein interaction; PPI) 자료와 같은 유전자 네트워크 자료를 통합 분석하는 방법들이 많이 제안되었다[6,7]. 이러한 방법들은 암과 같이 복잡한 질병들이 단일 유전자 보다는 여러 유전자의 발현 패턴에 의해 결정된다는 사실을 반영하기 위해, 유전자 네트워크 상의 유전자 모듈 또는 그래프 중심성(graph centrality)이 큰 유전자를 바이오마커(biomarker)로 식별하는 전략을 취했다.

† 교신 저자

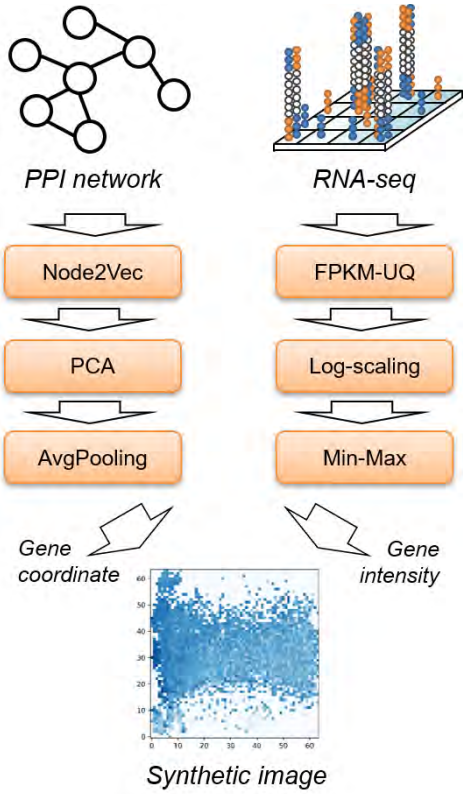
* 이 논문은 과학기술정보통신부와 한국연구재단의 방사선기술개발사업으로 지원을 연구 지원한 (2017M2A2A7A02020213)의 결과물입니다.

최근에는 네트워크의 구조 정보에 기반한 정점(node)의 분산 표현법(distributed representation)을 학습하여 암 예후 특이적 네트워크로부터 예후 예측에 관련된 유전자를 식별하는 방법이 제안되었다[8].

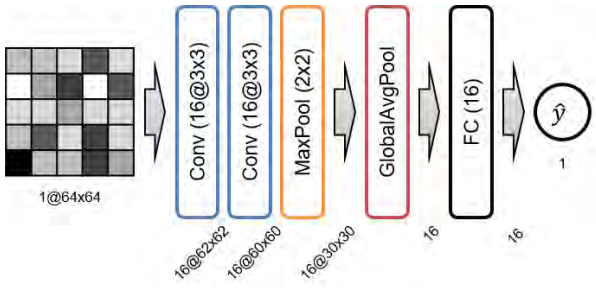
기존 연구들은 특징 선택(feature selection)과 분류 모델의 학습을 독립적으로 수행하여 차원의 저주 문제를 회피하였다. 이런 경우 분류 모델의 예측 정확도는 향상될 수 있으나, 전체 유전자 데이터에 숨겨진 생물학적 정보나 유전자들 간의 상호작용 관계를 분류 모델이 포착하기 어려운 문제가 일어나게 된다.

본 연구에서는 특징 선택 과정 없이 유전자 발현량 벡터의 전체 정보를 최대한 보존하여 향상된 예후 예측을 수행할 수 있는 학습 전략을 제안한다. 제안하는 전략은 크게 2 단계로 구성된다. 첫번째 단계는 고차원 유전자 벡터를 Node2Vec[9]을 이용하여 2 차원 이미지로 변환하는 것이고, 두번째 단계에서는 변환된 유전자 발현량 이미지 데이터를 합성곱 신경망(convolutional neural network; CNN)으로 학습 및 예후 예측을 수행한다. 유전자 발현량 벡터의 각 유전자들이 평면 위에 사상될 때 생물학적으로 상호작용이 있는 유전자들이 이웃할 수 있도록 유전자 네트워크 데이터와 Node2Vec 을 활용한다. Node2Vec 은 단백질-단백질 상호작용 네트워크 상에서 구조적 기능이 유사한 유전자들이 분산 표현 공간에서 이웃한 위치에 놓이도록 2 차원 벡터 값을 계산한다. 합성곱 신경망 모델은 유전자 발현량 이미지에 표현된 유전자 발현량 수치와 함께 Node2Vec 에 의해 내포된 유전자들 간의 상호작용 정보를 학습하여 예후 분류를 수행한다.

제안하는 전략의 성능 평가를 엄밀하게 하기 위해 이중 교차 검증(double cross-validation)[10] 및 무작위 탐색(random search)[11]을 이용하여 초매개변수 최적화(hyperparameter optimization) 및 예측 정확도 평가를 진행하였다.



(그림 1) Node2Vec 기반의 이미지 표현법



(그림 2) 암 예후 예측을 위한 합성곱 신경망 구조

2. 본론

2.1. 데이터 수집 및 정제

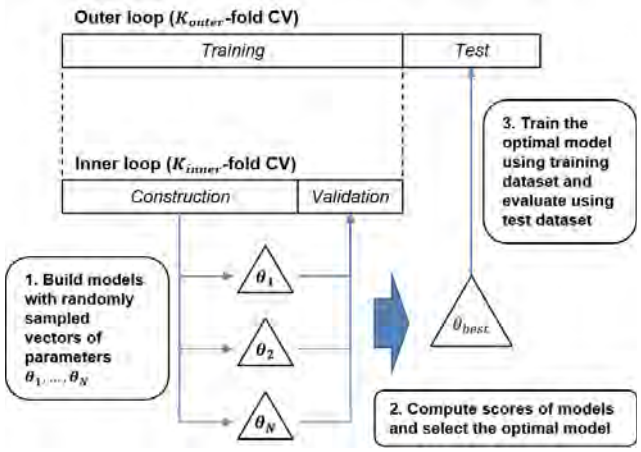
이차원 평면 상의 유전자 좌표를 구하기 위해 STRING 데이터베이스[12]로부터 단백질-단백질 상호작용 네트워크 데이터를 다운받았다. 다운받은 네트워크의 각 상호작용은 신뢰도 점수(confidence score)가 매겨져 있으며, 본 연구에서는 신뢰성이 높은 정보만 이용하기 위해 점수가 700 이하인 간선(edge)들을 모두 제거하였고, 그 결과 16,984 개의 유전자 정점과 420,875 개의 상호작용 간선을 수집하였다.

암환자의 유전자 발현량 데이터를 얻기 위해 R 패키지 중 하나인 TCGAbiolinks [13]를 이용하였다. TCGA 데이터베이스로부터 간암(LIHC), 폐암(LUAD), 췌장암(PAAD) 환자의 유전자 발현량 데이터 및 임상 자료(clinical data)를 다운로드하였다. 다운받은 유전자 발현량 데이터는 RNA-seq 기술로 측정된 FPKM-UQ 정규화 데이터이지만, 이상치(outlier)가 많고 분포가 심하게 치우쳐져 있어서 log 변환을 통해 데이터 분포의 치우침을 완화하였다[14]. 이후 샘플 간의 비교를 수행하기 위해 min-max 변환을 적용하여 0 과 1 사이의 크기로 변환하였다. 또한 단백질-단백질 상호작용 네트워크에 있는 유전자만 골라내는 작업을 수행하였고, 총 16,888 개의 유전자가 선택되었다.

<표 1> 유전자 발현량 데이터 요약

Cancer type	Poor prognosis	Good prognosis
간암(LIHC)	150	51
폐암(LUAD)	204	60
췌장암(PAAD)	92	8

예후 분류 문제를 수행하기 위해 임상 자료를 바탕으로 환자의 예후를 구분하였다. 임상 자료로부터 환자의 '사망여부' 및 '암 진단일로부터 마지막 관찰일까지 걸린 기간' 정보를 추출하였다. 암 확진 후 5 년 이내에 사망한 환자를 나쁜 예후 그룹으로, 5 년 이상 생존한 환자를 좋은 예후 그룹으로 구분하였으며, 각 암의 예후 그룹 별 정보를 표 1 에 나타냈다.



(그림 3) 이중 교차 검증 및 무작위 탐색

2.2 Node2Vec 기반의 이미지 표현법

그림 1 은 제안하는 이미지 표현 기법을 보여준다. Node2Vec 알고리즘은 단백질-단백질 상호작용 네트워크를 학습하고 각 유전자의 2 차원 벡터 표현을 계산한다. 계산된 벡터를 이용하여 각 유전자의 직교 좌표(orthogonal coordinate)를 계산하기 위해 주성분분석(principal component analysis; PCA)를 적용하였다. 연속형 값을 갖는 좌표를 이산화(discretization)하기 위해 직교축마다 비닝(equal-width binning)을 적용하여 64x64 크기 이미지의 픽셀에 유전자들을 배치하였다. 각 픽셀의 밝기(intensity)는 해당 픽셀에 배치된 여러 유전자들 발현량을 하나의 평균값으로 통합하여(Average pooling) 계산하였다.

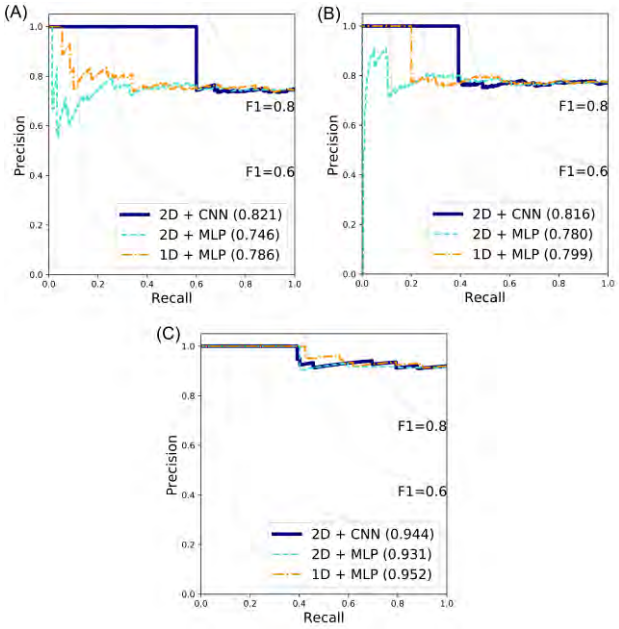
2.3 암 예후 예측을 위한 합성곱 신경망 구조

유전자 발현량 정보가 합성된 이미지(synthetic image)를 바탕으로 암환자의 예후를 예측하기 위해 합성곱 모델을 그림 2 와 같이 설계하였다. 2 개의 합성곱 계층은 16 개의 출력 필터, 3x3 크기의 핵(kernel), 1 칸의 스트라이드(stride), 그리고 패딩(padding)없는 연산을 수행한다. 1 개의 최대값풀링 계층은 2x2 크기의 핵, 2 칸의 스트라이드, 그리고 패딩없는 연산을 수행한다. 전역평균풀링 계층은 필터별로 평균을 계산하여 필터 개수와 동일한 크기의 1 차원 벡터를 출력한다. 이어서 완전 연결된(fully-connected) 크기 16 의 히든(hidden) 계층을 한번 거쳐서 최종 출력값인 예후 레이블을 반환한다. 모든 레이어 연산에서 활성화함수(activation function)는 배치정규화(batch normalization)와 ReLU 가 적용되었다. 완전 연결된 히든 계층에는 모델의 일반화 능력을 향상시키기 위해 L1 및 L2 정규화를 적용하였다. 손실함수는 cross-entropy 로 정의되었고, Adam 에 Nesterov 운동량이 적용된 Nadam 알고리즘[15]으로 모델 최적화를 수행하였다.

2.4 이중 교차 검증 및 무작위 탐색

모델의 예측 성능을 편향(bias)없이 평가하기 위해 이중 교차 검증을 수행하였다. 하나의 데이터집합을 가지고서 모델 선택(model selection)과 모델 평가(model assessment)를 각각의 교차 검증(cross validation)으로 수행할 경우 최종적인 모델 평가의 결과가 실제 그 모델이 갖는 예측 성능보다 더 높게 평가되는 문제가 있다[10]. 이를 방지하기 위해 이중 교차 검증은 훈련데이터(training)를 건설데이터(construction)와 검증데이터(validation)로 다시 나누는 것으로 교차 검증 안에서 다시 교차 검증을 수행한다(그림 3). 내부 교차 검증은 최적의 초매개변수 값을 결정하기 위한 과정이고, 외부 교차 검증은 내부 교차 검증에서 탐색된 매개변수를 가지고서 시험데이터(test)를 예측 및 성능 평가를 수행한다.

모델의 최적화를 효율적으로 수행하기 위해 무작위 탐색 전략을 사용하였다. 실험적으로 사용자가 직접 매개변수의 값을 바꾸어 가며 실험하는 것보다 무작위 탐색 전략을 사용하는 것이 시간 대비 효과적이라는 연구결과가 있다[11]



(그림 4) 암 예후 예측 모델의 예측 성능 비교를 위한 Precision-Recall 곡선. (A) 간암, (B) 폐암, (C) 췌장암; 2D+CNN 은 이미지 데이터에 CNN 을 적용한 경우, 2D+MLP 는 이미지 데이터를 1 차원 벡터로 변환하여 다층 퍼셉트론을 적용한 경우, 1D+MLP 는 유전자 발현량 벡터에 다층 퍼셉트론을 적용한 경우임. 괄호 안의 수치는 AUC 값임.

<표 2> 암 예후 예측 모델 성능표

Model	간암 (LIHC)	폐암 (LUAD)	췌장암 (PAAD)
2D+CNN	**0.821**	**0.816**	0.944
2D+MLP	0.746	0.780	0.931
1D+MLP	0.786	0.799	**0.952**

* 수치는 Precision-Recall 의 AUC 값임

3. 실험 결과

제안하는 유전자 발현량 이미지 표현법과 CNN 의

조합성능을 평가하기 위해 본래의 고차원 벡터를 이용하는 다층 퍼셉트론(Multi-Layer Perceptron) 모델과 예후 분류 정확도를 비교하였다. 그리고 합성된 발현량 이미지 데이터에 합성곱 신경망을 적용하는 것이 효과적임을 확인하기 위해 이차원 발현량 이미지를 1 차원 벡터로 평평하게 만들어 다층 퍼셉트론으로 예측하는 경우도 실험하였다.

각 모델의 분류 성능은 Precision-Recall 곡선 및 곡선 아래 면적(Area Under the Curve; AUC)으로 평가되었다(그림 4, 표 2). 첫번째 실험에서 간암과 폐암에 대해 제안하는 방법(2D+CNN)이 베이스라인 모델(baseline model)인 고차원 벡터를 그대로 사용하는 다층 퍼셉트론 모델(1D+MLP)보다 0.017-0.035 정도 향상된 예측 정확도를 보였다. 두번째 실험에서는 모든 암종에 대하여 다층 퍼셉트론을 적용한 경우(2D+MLP)보다 합성곱 신경망을 사용하는 것이 0.013-0.075 정도 더 높은 예측 성능을 보여줄 수 있다는 것을 확인하였다.

4. 분석 및 결론

본 연구에서는 향상된 예후 예측 정확도를 보여줄 수 있는 유전체 데이터에 특화된 분류 모델을 제안하였다. 제안하는 방법을 평가하기 위해 간암, 폐암, 그리고 췌장암의 예후 예측 정확도를 측정하였고, 고차원 벡터를 입력 받는 다층 퍼셉트론 모델보다 더 정확하게 예측할 수 있는 것을 확인하였다.

제안하는 방법은 개선의 여지가 여럿 있다. 본 연구에서는 Node2Vec 을 이용하여 유전자 좌표를 구하였으나, 이를 대신할 수 있는 알고리즘으로 DeepWalk, 또는 LINE 등이 개발되어 있다[16]. 각각의 알고리즘으로 비교 실험을 하여 가장 효과적인 방법이 무엇인지 연구할 필요가 있다. 합성곱 신경망 모델에 관해서도 개선할 부분이 있다. 본 연구에서 사용한 데이터는 크기가 몹시 작아, 얕은 합성곱 신경망 모델을 사용하였다. 신경망이 깊어지고 많은 샘플로 학습되면 예측 정확도가 더욱 높아질 수 있다고 기대되기 때문에, 향후 연구에서는 전이 학습(transfer learning) [17] 같은 전략을 활용하여, 샘플이 충분히 많은 암종 데이터로부터 깊은 신경망을 학습시키고 작은 크기의 암종 데이터에 응용하여 효과적으로 분석할 수 있는 전략을 연구할 계획이다.

참고문헌

[1] Min, Seonwoo, Byunghan Lee, and Sungroh Yoon. "Deep learning in bioinformatics." Briefings in bioinformatics 18.5 (2017): 851-869.

[2] Martínez-Ballesteros, María, et al. "Machine learning techniques to discover genes with potential prognosis role in Alzheimer's disease using different biological sources." Information Fusion 36 (2017): 114-129.

[3] 보건복지부, "암등록통계 (국가승인통계 117044 호)." (2018).

[4] Sotiriou, Christos, et al. "Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis." Journal of the National Cancer Institute 98.4 (2006): 262-272.

[5] Clarke, Robert, et al. "The properties of high-dimensional data spaces: implications for exploring gene and protein expression data." Nature reviews cancer 8.1 (2008): 37.

[6] Choi, Jonghwan, et al. "Improved prediction of breast cancer outcome by identifying heterogeneous biomarkers." Bioinformatics 33.22 (2017): 3619-3626.

[7] Martinez-Ledesma, Emmanuel, Roeland GW Verhaak, and Victor Treviño. "Identification of a multi-cancer gene expression biomarker for cancer clinical outcomes using a network-based algorithm." Scientific reports 5 (2015): 11966.

[8] Choi, Jonghwan, et al. "G2Vec: Distributed gene representations for identification of cancer prognostic genes." Scientific reports 8.1 (2018): 13729.

[9] 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.

[10] Varma, Sudhir, and Richard Simon. "Bias in error estimation when using cross-validation for model selection." BMC bioinformatics 7.1 (2006): 91.

[11] Bergstra, James, and Yoshua Bengio. "Random search for hyper-parameter optimization." Journal of Machine Learning Research 13.Feb (2012): 281-305.

[12] Szklarczyk, Damian, et al. "STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets." Nucleic acids research 47.D1 (2018): D607-D613.

[13] Colaprico, Antonio, et al. "TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data." Nucleic acids research 44.8 (2015): e71-e71.

[14] Danielsson, Frida, et al. "Assessing the consistency of public human tissue RNA-seq data sets." Briefings in Bioinformatics 16.6 (2015): 941-949.

[15] Dozat, Timothy. "Incorporating nesterov momentum into adam." (2016).

[16] Qiu, Jiezhong, et al. "Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec." Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. ACM, 2018.

[17] Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." IEEE Transactions on knowledge and data engineering 22.10 (2010): 1345-1359.