KIPS Trans. Softw. and Data Eng.
Vol.8, No.10 pp.397~402 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2019.8.10.397

A Node2Vec-Based Gene Expression Image Representation Method for Effectively Predicting Cancer Prognosis

Jonghwan Choi† · Sanghyun Park††

ABSTRACT

Accurately predicting cancer prognosis to provide appropriate treatment strategies for patients is one of the critical challenges in bioinformatics. Many researches have suggested machine learning models to predict patients' outcomes based on their gene expression data. Gene expression data is high-dimensional numerical data containing about 17,000 genes, so traditional researches used feature selection or dimensionality reduction approaches to elevate the performance of prognostic prediction models. These approaches, however, have an issue of making it difficult for the predictive models to grasp any biological interaction between the selected genes because feature selection and model training stages are performed independently. In this paper, we propose a novel two-dimensional image formatting approach for gene expression data to achieve feature selection and prognostic prediction effectively. Node2Vec is exploited to integrate biological interaction network and gene expression data and a convolutional neural network learns the integrated two-dimensional gene expression image data and predicts cancer prognosis. We evaluated our proposed model through double cross-validation and confirmed superior prognostic prediction accuracy to traditional machine learning models based on raw gene expression data. As our proposed approach is able to improve prediction models without loss of information caused by feature selection steps, we expect this will contribute to development of personalized medicine.

Keywords : Bioinformatics, Gene Expression, Node2Vec, Cancer Prognostic Prediction, Personalized Medicine

암 예후를 효과적으로 예측하기 위한 Node2Vec 기반의 유전자 발현량 이미지 표현기법

최 종 환† · 박 상 현††

요 약

암 환자에게 적절한 치료계획을 제공하기 위해 암의 진행양상 또는 환자의 생존 기간 등에 해당하는 환자의 예후를 정확히 예측하는 것은 생물정보학 분야에서 다루는 중요한 도전 과제 중 하나이다. 많은 연구에서 암 환자의 유전자 발현량 데이터를 이용하여 환자의 예후를 예측하는 기계학습 모델들이 많이 제안되어 오고 있다. 유전자 발현량 데이터는 약 17,000개의 유전자에 대한 수치값을 갖는 고차원의 수치형 자료이기에, 기존의 연구들은 특징 선택 또는 차원 축소 전략을 이용하여 예측 모델의 성능 향상을 도모하였다. 그러나 이러한 접근법은 특징 선택과 예측 모델의 훈련이 분리되어 있어서, 기계학습 모델은 선별된 유전자들이 생물학적으로 어떤 관계가 있는지 알기가 어렵다. 본 연구에서는 유전자 발현량 데이터를 이미지 형태로 변환하여 예후 예측이 효과적으로 특징 선택 및 예후 예측을 수행할 수 있는 기법을 제안한다. 유전자들 사이의 생물학적 상호작용 관계를 유전자 발현량 데이터에 통합하기 위해 Node2Vec을 활용하였으며, 2차원 이미지로 표현된 발현량 데이터를 효과적으로 학습할 수 있도록 합성곱 신경망 모델을 사용하였다. 제안하는 모델의 성능은 이중 교차검증을 통해 평가되었고, 유전자 발현량 데이터를 그대로 이용하는 기계학습 모델보다 우월한 예후 예측 정확도를 가지는 것이 확인되었다. Node2Vec을 이용한 유전자 발현량의 새로운 이미지 표현법은 특징 선택으로 인한 정보의 손실이 없어 예측 모델의 성능을 높일 수 있으며, 이러한 접근법이 개인 맞춤형 의학의 발전에 이바지할 것으로 기대한다.

키워드 : 생물정보학, 유전자 발현량, Node2Vec, 암 예후 예측, 맞춤형 의학

1. 서 론

데이터 공학 기술의 발전은 4차산업혁명에 핵심 기술로 활용되고 있으며, 특히 스마트시티(smart city), 맞춤형 의학(personalized medicine) 등 현대인들에게 이로운 서비스를 제공할 수 있게 하였다. 생물정보학(bioinformatics) 분야에서는 맞춤형 의학 서비스의 도래를 위해 기계학습(machine learning) 및 심층 학습(deep learning) 모델을 활용한 질병 연구가 많이 이루어졌고[1], 특히 암 환자의 생존율을 높이기 위한 암 예후(cancer prognosis) 연구가 활발히 이루어졌다[2]. 암 환자의 예후란, 환자의 종양(tumor)이 다른 기관으로 전이(metastasis)될지, 또는 환자의 생명이 곧 다할지 등 환자의 전망(outcome)을 예측하는 것을 말한다. 보건복지부의 2018년 12월에 발표된 국가암등록통계 자료에 의하면, 국내 폐암, 간암, 췌장암 환자의 5년 생존율이 각각 28.2%, 34.6%,

※ 이 논문은 국토교통부의 스마트시티 혁신인재육성사업으로 지원되었습니다.
※ 이 논문은 2019년도 한국정보처리학회 춘계학술발표대회에서 '암 유전체 데이터를 효과적으로 학습하기 위한 Node2Vec 기반의 새로운 2차원 이미지 표현기법'의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 연세대학교 컴퓨터과학과 박사과정
†† 종신회원 : 연세대학교 컴퓨터과학과 교수
Manuscript Received : July 5, 2019
Accepted : July 25, 2019
* Corresponding Author : Sanghyun Park(sanghyun@yonsei.ac.kr)

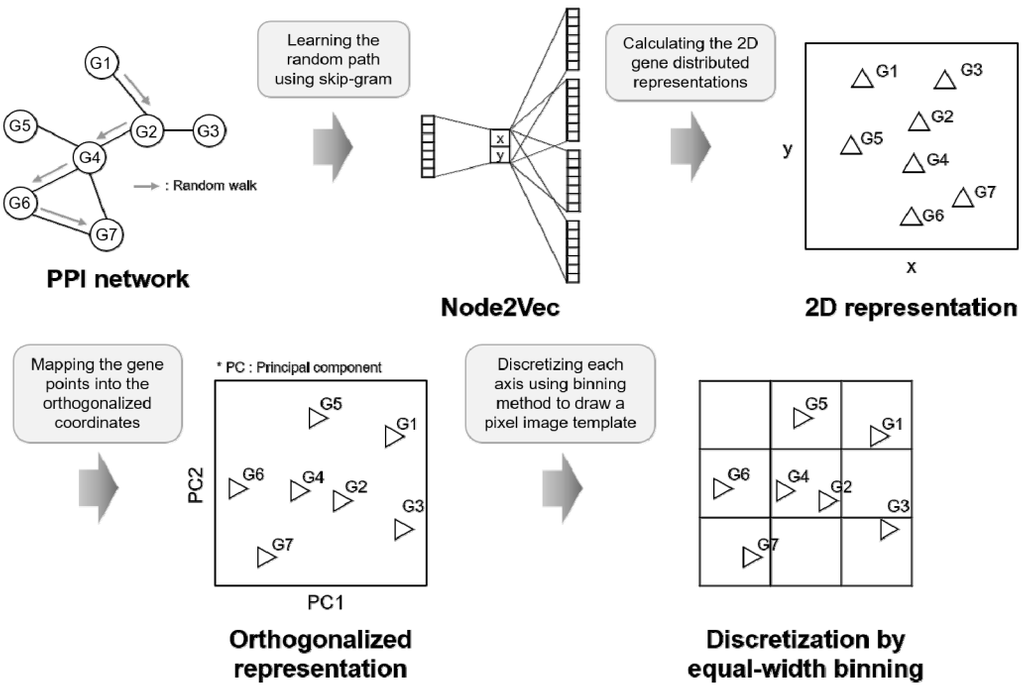

Fig. 1. Procedure of Calculating PPI Network-Based Gene Coordinates Via Node2Vec

11.4%로 매우 나쁜 예후를 가지는 것으로 보고되었으며[3], 이들을 포함한 여러 암종의 생존율을 높이기 위해, 개인 환자 정보에 기반한 예후 예측 모델의 연구 및 개발이 필요한 실정이다.

2. 선행 연구

암 환자의 예후를 예측하기 위해 널리 이용되는 대표적인 유전체 자료에는 유전자 발현량 데이터(gene expression data)가 있다[4]. 유전자 발현량 데이터는 단백질 정보를 암호화하고 있는 유전자(protein-coding gene)의 활성화 정도를 나타내는 수치형 자료(numerical data)이다. 사람은 2만여 개의 유전자를 가지고 있으며, 암 예후에 관련된 기계학습 모델들은 암 환자의 유전자 발현량 벡터를 입력받아 환자의 예후가 좋은지 나쁜지를 예측한다[2]. 그러나, 기계학습 모델이 큰 훈련 집합(training set)을 요구하는 반면에 유전자 발현량에 있는 환자의 수는 극히 한정적이고, 또 특징 개수가 2만여 개나 있어서 기계학습의 성능을 충분히 활용하기가 어렵다[5]. 이를 위해 기존의 많은 연구에서는 특징 선택(feature selection) 또는 차원 축소 기법(dimensional reduction)들을 적용하여 유전자의 수를 줄이는 접근법이 널리 사용되었다[6].

유전자 발현량 데이터의 효과적인 차원 축소를 위해 단백질-단백질 상호작용(protein-protein interaction; PPI) 네트워크와 같은 유전자들의 생물학적 관계 정보를 활용하는 분석 기법들이 여럿 제안되었다[7, 8]. PPI 네트워크에서 단백질 및 유전자에 해당하는 정점(vertex)의 그래프 중심성(graph centrality)을 계산하여 중요한 유전자를 식별하거나[7], 유전자 모듈(module)을 탐색하여 예후와 관련된 유전자 집합을 찾는 접근법들이 많이 개발되었다[8]. 최근에는 PPI 네트워크의 지역적 및 전역적 구조 정보를 학습한 정점의 분산 표현(distributed representation)을 계산하여, 이로부터 예후 바이오마커(prognostic biomarker)를 선택하는 방법이 제안되었다[9].

특징 선택을 이용한 기존의 전략들은 유전자 발현량 데이터가 가지고 있는 차원의 저주(curse of dimensionality) 문제를 완화하고 암 예후 예측 정확도의 향상을 보여 주었다. 하지만 기존 모델의 대부분에서는 차원 축소 단계와 예측 모델 훈련 단계가 분리되어 있으며, 선별된 일부 유전자들에 대해서만 기계학습 모델의 훈련이 이루어진다. 분리된 단계는 기계학습 모델이 유전자 발현량 데이터 전반에 내재하는 생물학 및 의학에서 알고자 하는 새로운 정보를 포착해줄 수 없으며, 또한 주어진 유전자들이 생물학적 관련성을 갖는지 알기가 어렵다.

위의 문제를 해결하기 위해, 본 연구에서는 특징 추출과 모델 훈련이 동시에 효과적으로 수행될 수 있도록 두 가지 심층학습 모델을 이용한 예후 예측 기법을 소개한다. 제안하는 기법은 크게 2단계로, 먼저 Node2Vec[10] 및 PPI 네트워크를 이용하여 유전자 발현량 데이터를 이미지 형태로 변환하고, 다음으로 합성곱 신경망 모델(convolutional neural network)을 이용하여 발현량 이미지 기반의 예후 예측을 진행한다. 이미지 형태로 변환하는 목적은 단순히 값만 나열되어있는 유전자 발현량 데이터에 유전자 간의 생물학적 상호작용 관계 정보를 통합하기 위함이며, 합성곱 연산은 통합된 상호작용 정보를 이용하여 유전자 발현량 데이터가 효과적으로 학습될 수 있도록 만든다.

제안하는 모델의 예후 예측 성능을 엄격하게 평가하기 위

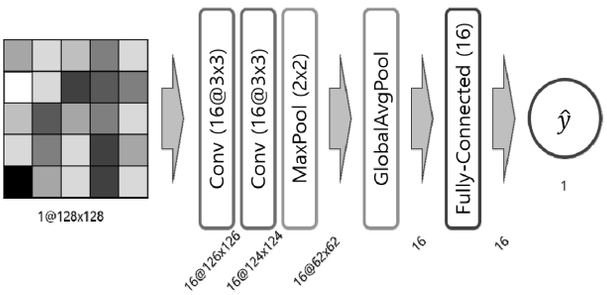

Fig. 2. An Architecture of Convolutional Neural Network for Predicting Cancer Outcomes

해 이중 교차검증(double cross-validation)[11]을 수행하였으며, 훈련 데이터에 맞추어 자동으로 하이퍼 파라미터 최적화(hyper-parameter optimization)가 이루어지도록 무작위 탐색(random search)[12] 전략을 병행하였다. 상기 검증 방법을 통해 제안하는 모델이 폐암, 간암, 췌장암을 포함한 7가지의 암종에 대하여 유전자 발현량 데이터를 그대로 입력받는 다층 퍼셉트론(multi-layer perceptron) 모델보다 우수한 예후 예측 정확도를 보여 줌을 확인하였다. 또한, 유전자 발현량 이미지 데이터에 합성곱이 없는 예측 모델을 적용한 경우와도 비교하여, 이미지에 내포된 유전자 상호작용 정보가 예측 정확도 향상에 기여하고 있음을 확인하였다.

3. 제안 모델

3.1 Node2Vec 기반의 이미지 표현법

고차원의 유전자 발현량 벡터를 2차원 이미지로 표현할 때, 생물학적 상호작용 관계가 있는 유전자들의 좌표가 인접하도록 만들기 위해 Node2Vec 및 PPI 네트워크를 이용하였다.

Node2Vec은 자연어처리에서 활용되는 Word2Vec 모델 중 하나인 skip-gram 모델을 그래프 데이터에 적용하여 정점에 대한 분산 표현을 구하는 심층학습 모델이다[10]. Node2Vec은 무작위행보(random walk) 알고리즘을 통해 그래프에서 얻은 경로(path)를 문장(sentence)으로, 경로 위의 정점들을 단어(word)로 인식하여 그래프 구조 정보가 반영된 정점들의 분산 표현을 계산한다. 이러한 Node2Vec을 PPI 네트워크에 적용하여 유전자들의 상호작용 정보가 반영된 유전자 분산 표현을 계산하였으며, 분산 표현의 크기를 2차원으로 제한하여 얻어진 두 개의 값을 각각 이미지의 x, y 좌표로 사용하였다. Node2Vec으로 계산한 유전자의 좌표는 주성분 분석(principal coordinate analysis)을 통해 직교화(orthogonization) 되었고, 균등 너비 이산화(equal width binning)을 적용하여 수치형 좌표를 픽셀(pixel) 이미지의 정수형 좌표로 변환하였다(Fig. 1).

PPI 네트워크에 기반한 유전자 픽셀 좌표를 계산 후, 유전자 발현량 데이터를 이용하여 각 픽셀의 세기(intensity)를 부여하였다. 발현량 값이 0에 가까울수록 해당 픽셀은 어둡게, 1에 가까울수록 픽셀을 밝게 만든다. 두 개 이상의 유전자가 같은 픽셀에 놓이는 경우, 해당 유전자들의 발현량 값의 평균으로 픽셀의 세기를 계산하였다.

3.2 암 예후 예측을 위한 합성곱 신경망 모델

각 암 환자의 유전자 발현량 데이터와 PPI 네트워크가 합성된 이미지를 바탕으로 환자의 예후 그룹을 예측하기 위해 이미지 데이터 학습에 효과적인 합성곱 신경망 모델을 사용하였다(Fig. 2). 본 연구에서 사용한 합성곱 신경망 모델은 2개의 합성곱 계층(convolutional layer)과 1개의 최대값 풀링 계층(max pooling layer)를 이용하여 이미지 데이터로부터 특징을 추출하고, 전역 평균 풀링(global average pooling) 단계를 통해 특징 벡터로 변환 및 1개의 완전 연결 계층(fully connected layer)을 통해 암 환자의 예후 그룹이 좋은 그룹인지 나쁜 그룹인지를 예측하였다.

사용된 합성곱 신경망 모델의 세부 디자인은 다음과 같다. 두 개의 합성곱 계층은 16개의 출력 필터(filter), 3x3 크기의 핵(kernel), 1칸의 스트라이드(stride), 그리고 패딩(padding) 없는 연산을 수행한다. 1개의 최대값 풀링 계층은 2x2 크기의 핵, 2칸의 스트라이드, 그리고 패딩없는 연산을 수행한다. 모든 계층에서 활성화 함수(activation function)로 ReLU 함수가 적용되었으며, 원활한 학습을 도모하기 위해 배치 정규화(batch normalization)를 적용하였다. 완전 연결 계층에 대해서는 모델의 일반화 능력 향상을 위한 L1 및 L2 정규화가 적용되었다. 이진 분류 작업을 위한 손실함수(loss function)로 크로스 엔트로피(cross entropy) 함수를 사용하였으며, 신경망 모델의 빠른 훈련을 위해 Adam 알고리즘에 Nesterov 운동량이 적용된 Nadam 알고리즘[13]을 이용하였다.

4. 실험 방법 및 결과

본 연구에서는 두 가지 실험을 수행하였다. 첫 번째 실험에서는 유전자 발현량 데이터 및 PPI 네트워크 데이터를 합성한 이미지에 대한 최적의 이미지 크기를 탐색하였고, 두 번째 실험에서는 7개의 암종 각각에 대하여 제안하는 모델의 예후 예측 성능을 평가하였다. 두 실험에서 예후 예측 정확도는 나쁜 예후를 얼마나 정확히 예측하였는가를 측도할 수 있는 Precision-Recall 곡선 아래의 면적(area under curve; AUC)으로 계산되었다.

4.1 데이터 수집 및 정제

본 연구에서는 세계적으로 가장 많은 암 환자의 유전체 정보를 가지고 있는 TCGA 공개 데이터베이스[14]로부터 7개의 암종, 신장암(BLCA), 자궁경부암(CESC), 뇌종양(LGG), 간암(LIHC), 폐암(LUAD), 난소암(OV), 췌장암(PAAD)에 대한 환자들의 임상 정보(clinical data) 및 유전자 발현량 데이터 집합을 수집하였다. 데이터 다운로드는 R 패키지 중 하나인 TCGAbiolinks[15]를 통해 이루어졌다.

유전자 발현량 데이터는 16,888개의 유전자를 갖는 수치형 자료로써 두 번의 전처리가 이루어졌다. 확보한 발현량 데이터는 RNA-seq 기반의 FPKM-UQ 정규화가 적용된 것이며, 이러한 데이터는 이상치(outlier)가 많고, 발현량 값의 분포가 심하게 치우쳐져 있어서[16], 로그 변환(log transformation)을 적용하여 치우침의 문제를 완화하였다. 이어서 변환된 데

Table 1. Prognostic Groups Information of 7 Cancer Types

Cancer types	Number of patients	Poor prognosis	Good prognosis
BLCA	233	186	47
CESC	108	64	44
LGG	177	106	71
LIHC	201	150	51
LUAD	264	204	60
OV	271	194	77
PAAD	100	92	8

이터를 정규화하기 위해 환자별로 min-max 변환을 적용하여 발현량의 크기를 0과 1 사이로 조정하였다.

유전자 발현량 데이터와 함께 TCGA 데이터베이스로부터 내려받은 임상 자료를 바탕으로 암 환자들을 두 개의 예후 그룹으로 구분하였다. 임상 자료로부터 3개의 예후 정보, 생존 여부(vital status), 암 진단일로부터 사망까지의 생존 기간(days to death), 암 진단일로부터 마지막 관찰까지의 생존 기간(days to last follow up)을 수집하였으며, 암 진단일로부터 5년 이내에 사망한 암 환자를 나쁜 예후(poor prognosis) 그룹으로, 5년 이상 생존하고 있는 환자를 좋은 예후(good prognosis)로 분류하였다(Table 1).

다음으로 유전자 발현량 벡터를 2차원 이미지로 변환하는 데 필요한 PPI 네트워크 자료를 STRING 공개 데이터베이스[17]로부터 내려받았다. 내려받은 네트워크는 무방향성 그래프(undirected graph)이며, 각 간선(edge)에는 해당 상호작용에 대한 신뢰도 점수(confidence score)가 부여되어 있다. 본 연구에서는 신뢰도 점수가 700 이상인 간선만 골라냈고, 그 결과 420,875개의 상호작용 간선이 선택되었다.

4.2 이중 교차검증 및 무작위 탐색

모델의 예측 성능을 엄밀하게 평가하기 위해 K겹 교차 검증(K-fold cross-validation) 및 이중 교차검증 방법을 사용하였다(Fig. 3). K겹 교차검증은 하나의 데이터 집합을 K개의 부분집합으로 나누고, 1개의 부분집합을 시험 집합(test set)으로, 나머지 K-1개를 훈련 집합으로 할당하여 모델 학습 및 평가를 진행한다. 이때 K개의 부분집합이 모두 한 번씩 번갈아 가며 시험 집합으로 사용되며, 이로부터 계산된 K개의 예측 정확도의 평균값을 모델의 최종 예측 정확도로 평가된다.

두 개 이상의 모델을 비교해야 하는 경우, 각 모델에 대한 파라미터 조정(parameter tuning)과 모델 평가(model assessment) 단계가 서로 독립적인 데이터 집합으로 수행되어야 한다. 독

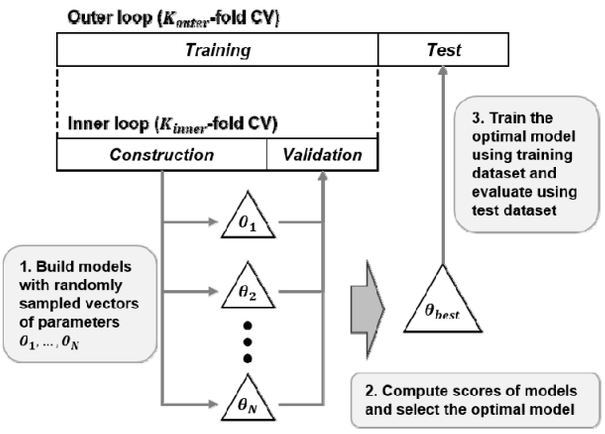

Fig. 3. Double Cross-validation and Random Search

립적이지 않은 데이터를 가지고서 모델의 최적 파라미터를 찾고, 그 파라미터가 적용된 모델을 같은 데이터 집합으로 평가하면, 성능 평가 결과는 실제 예측 성능보다 고평가되는 위험이 있다[11]. 이러한 편향(bias)을 방지하기 위해 이중 교차검증은 훈련 집합에 K겹 교차검증을 적용하여 건설 집합(construction set)과 검증 집합(validation set)을 통해 파라미터 조정을 수행하고, 최적 모델 선택에 전혀 이용되지 않았던 시험 집합으로 모델 평가를 수행한다.

이중 교차검증의 파라미터 조정 단계에서 하이퍼 파라미터들의 가능한 모든 조합을 조사하는 것은 현실적으로 어려운 일이다. 모든 조합을 조사하는 문제를 회피하면서 적은 횟수의 실험으로 최적에 근사한 파라미터 조합을 찾기 위해, 본 연구에서는 무작위 탐색(random search) 전략을 취하였다. 무작위 탐색이란 사용자가 정의한 파라미터 공간(parameter space) 내에서 지정된 개수만큼 파라미터 조합들을 무작위로 골라내고, 이들 중에서 최적의 조합을 선택하는 모델 최적화 방법이다. 이 탐색 방법이 많은 연구에서 사용자가 직접 파라미터를 조정하는 경우보다 시간 대비 효과적이라는 보고가 있다[12]. 본 연구에서는 총 3개의 예측 모델을 실험하였으며, Table 2는 각 모델이 갖는 하이퍼 파라미터들에 대하여 지정된 상한(upper bound), 하한(lower bound), 확률분포(probability distribution)를 보여준다.

4.3 실험 환경

실험에 사용된 CPU는 Intel Xeon E5-2640이고 DRAM은 64GB로 구성하였다. GPU로 Nvidia Geforce RTX 2080을 활용하여 딥러닝 모델의 훈련을 가속하였다. 실험 소스코드는

Table 2. Hyper-parameter Distributions for Random Search

Model	Hyper-parameter	Lower bound	Upper bound	Probability distribution
Convolutional Neural Network	learning rate	1E-03	1E-01	continuous log-uniform
	l1 regularization strength	1E-05	1E-03	continuous log-uniform
	l2 regularization strength	1E-05	1E-03	continuous log-uniform
Multi-layer Perceptron	learning rate	1E-03	1E-01	continuous log-uniform
	hidden layer width	{8, 32, 128}		discrete uniform
	hidden layers depth	{1, 2}		discrete uniform
	l2 regularization strength	1E-05	1E-03	continuous log-uniform

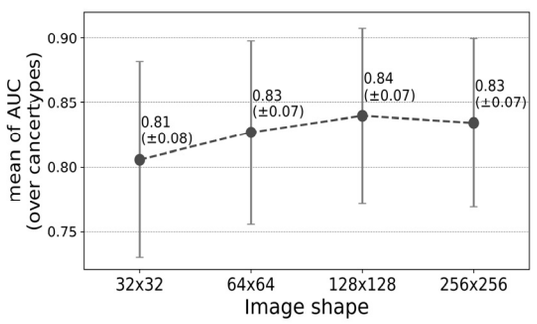

Fig. 4. Prediction Performance on Four Gene Expression Image Shapes

모두 Python 3로 작성되었으며, Node2Vec 및 합성곱 신경망 모델은 TensorFlow 1.12 및 Keras로 구현되었고, 이중 교차 검증 및 무작위 탐색은 Scikit-learn 라이브러리에서 제공하는 것을 활용하였다.

4.4 최적의 이미지 크기 탐색

제안하는 이미지 표현기법은 고차원의 유전자 발현량 벡터를 PPI 네트워크에 기반한 2차원 픽셀 이미지로 사상시키며, 이를 통해 유전자들 사이의 생물학적 상호작용 정보를 융합시킨 유전자 발현량 이미지를 만든다. 유전자 발현량 이미지의 크기가 작을수록 한 픽셀에 중첩되는 유전자 수가 많아져서 입력값의 차원을 효과적으로 줄일 수 있지만, 정보가 크게 손실될 수 있는 우려가 발생한다. 반면에 이미지 크기가 너무 크면 유전자들이 멀리 위치하게 되어 합성곱 연산이 활용되더라도 네트워크의 상호작용 정보가 무의미할 수 있다.

최적의 이미지 크기를 찾기 위해 4가지의 크기(32x32, 64x64, 128x128, 256x256)에 대하여 모델의 예측 성능을 평가 및 비교하였다. 각 이미지 크기 별로 7가지 암종에 대한 예후 예측 정확도를 측정하였고, 각 크기에 대한 평균과 표준편차를 계산하였다. Fig. 4는 각 크기의 평균과 표준편차를 막대 그래프로 나타낸 그래프이며, 128x128의 경우가 가장 높은 기대 성능을 보였다. 그리고 이미지 크기가 너무 작거나 너무 커지면 예후 예측 성능이 저조해지는 것도 확인하였다.

4.5 제안 모델 성능 평가

제안하는 기법의 핵심적인 특징은 두 가지로, 기존의 유전자 발현량 벡터를 이미지로 표현하는 것과 표현된 이미지를 합성곱 신경망을 통해 효과적으로 학습하는 것이다. 첫 번째 특징이 예후 예측 정확도 향상에 공헌하는지 확인하기 위해 유전자 발현량 벡터를 그대로 학습하는 다층 퍼셉트론의 AUC값과 비교하였다. 이어서 합성곱 신경망이 이미지 형태의 유전자 발현량 데이터를 효과적으로 활용한다는 두 번째 특징을 검증하기 위해, 이미지로 표현된 발현량 데이터를 입력받는 다층 퍼셉트론 모델과 비교하였다. 이미지로 변환되는 과정에서 PPI 네트워크의 정보에 따라 상호작용이 있는 유전자들의 발현량 데이터가 유의미하게 뭉쳐졌기에, 발현량 벡터에 대한 다층 퍼셉트론과 이미지에 대한 다층 퍼셉트론은 다른 결과를 보여 준다. Fig. 5는 7개의 암종에 대하여 예후 예측 정확도를 측정한 결과를 보여 준다. 제안하는 모델(image+CNN)이 유전자 발현량 데이터를 그대로 입력받는 경우(vector+MLP)보다 더 높은 예측 정확도를 보이는 것을 확인하였고, 이미지로 표현된 환자의 발현량 데이터를 다층 퍼셉트론으로 학습하는 경우(image+MLP)보다 합성곱 신경망을 쓰는 것이 더 효과적이라는 것 또한 확인할 수 있었다.

5. 결 론

본 연구에서는 암 환자의 예후를 보다 정확하게 예측하기 위해 고차원의 유전자 발현량 데이터를 효과적으로 학습할 수 있는 새로운 이미지 표현기법 및 적합한 심층학습 모델을 제안하였다. 제안하는 표현 및 예측 전략은 7개의 암종(신장암, 자궁경부암, 뇌종양, 간암, 폐암, 난소암, 췌장암)에 대하여 기존의 기계학습 모델보다 향상된 예후 예측 정확도를 보여 주어, 제안 모델이 유전체 정보에 기반한 개인 맞춤형 치료전략을 실현하는데 공헌할 것으로 기대한다.

제안하는 방법은 더욱 향상될 여지가 있다. PPI 네트워크 데이터와 유전자 발현량 데이터를 융합하기 위해 Node2Vec을 활용하였다. 하지만 DeepWalk 또는 LINE 등 다른 여러 가지 정점 분산 표현기법들이 개발되어 있기에[18], 이들을 통해 더 효과적인 이미지 좌표를 계산할 수 있는지를 비교 분석해볼 필요가 있다. 다른 개선 사항으로, 본 연구에서는 간단한 합성곱 신경망 구조를 사용하였기에, 더 깊고 복잡한 구조를 사용하여 예후 예측 정확도를 높일 수 있는지 또한 연구해볼 필요가 있다.

향후 연구에서는 상기된 개선 사항에 관한 조사뿐만 아니

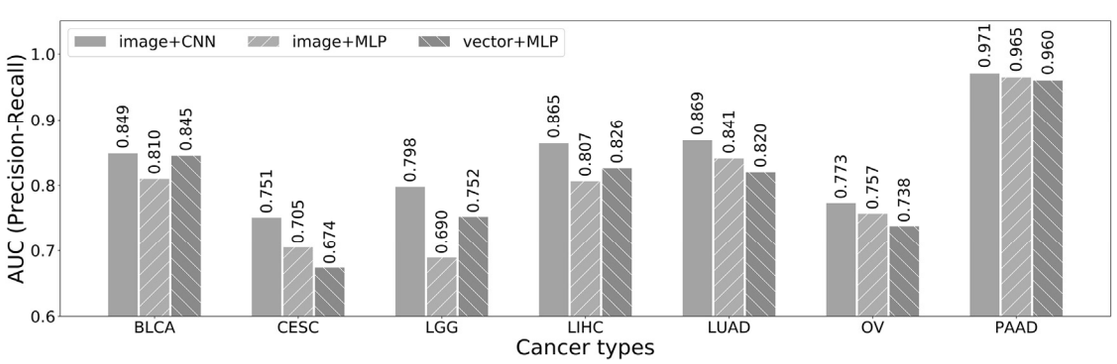

Fig 5. Evaluation of prognostic prediction accuracy between the proposed method and baseline models

라, 암 환자의 유전자 발현량 데이터에 다른 유전체 데이터들을 통합한 오믹스(omics) 데이터를 학습할 수 있는 모델로 발전시켜 볼 계획이다.

References

[1] S. W. Min, B. G. Lee, and S. R. Yoon, "Deep learning in bioinformatics," *Briefings in Bioinformatics*, Vol.18, No.5, pp.851-869, 2017.

[2] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, Vol.13, pp.8-17, 2015.

[3] Ministry of Health and Welfare, Republic Korea, "National Cancer Statistics in 2016." 2018.

[4] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, C. Desmedt, D. Larsimont, F. Cardoso, H. Peterse, D. Nuyten, M. Buyse, M. J. Van de Vijver, J. Bergh, M. Piccart, and M. Delorenzi, "Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis," *Journal of the National Cancer Institute*, Vol.98, No.4, pp.262-272, 2006.

[5] R. Clarke, H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, and Y. Wang, "The properties of high-dimensional data spaces: implications for exploring gene and protein expression data," *Nature Reviews Cancer*, Vol.8, No.1, pp.37, 2008.

[6] L. Wang, Y. Wang, and Q. Chang, ""Feature selection methods for big data bioinformatics: A survey from the search perspective," *Methods*, Vol.111, pp.21-31, 2016.

[7] J. Choi, S. Park, Y. Yoon, and J. Ahn, "Improved prediction of breast cancer outcome by identifying heterogeneous biomarkers," *Bioinformatics*, Vol.33, No.22, pp.3619-3626, 2017.

[8] E. Martinez-Ledesma, R. G. W. Verhaak, and V. Treviño, "Identification of a multi-cancer gene expression biomarker for cancer clinical outcomes using a network-based algorithm," *Scientific Reports*, Vol.5, pp.11966, 2015.

[9] J. Choi, I. Oh, S. Seo, and J. Ahn, "G2Vec: Distributed gene representations for identification of cancer prognostic genes," *Scientific Reports*, Vol.8, No.1, pp.13729, 2018.

[10] A. Grover, and J. Leskovec, "Node2vec: Scalable feature learning for networks," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2016.

[11] S. Varma, and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC* Bioinformatics, Vol.7, No.1, pp.91, 2006.

[12] J. Bergstra, and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, Vol.13(Feb.), pp.281-305, 2012.

[13] T. Dozat, "Incorporating nesterov momentum into adam," 2016.

[14] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge," *Contemporary Oncology*, Vol.19, No.1A, pp.A68, 2015.

[15] A. Colaprico, T. C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, T. S. Sabedot, T. M. Malta, S. M. Pagnotta, I. Castiglioni, M. Ceccarelli, G. Bontempi, and H. Noushmehr, "TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data," *Nucleic Acids Research*, Vol.44, No.8, pp.e71-e71, 2015.

[16] F. Danielsson, T. James, D. Gomez-Cabrero, and M. Huss, "Assessing the consistency of public human tissue RNA-seq data sets," *Briefings in Bioinformatics*, Vol.16, No.6, pp.941-949, 2015.

[17] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, and C. Mering, "STRING v11: protein - protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Research*, Vol.47, No.D1, pp.D607-D613, 2018.

[18] J. Qiu, Y. Dong, H. Ma, J. Li, and K. Wang, "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.

최 종 환

https://orcid.org/0000-0002-8429-4135
e-mail : mathcombio@yonsei.ac.kr
2016년 인천대학교 수학과(이학사)
2018년 인천대학교 컴퓨터공학과 (공학석사)
2019년~현 재 연세대학교 컴퓨터과학과 박사과정
관심분야 : 바이오인포매틱스, 기계학습, 심층학습, 데이터마이닝

박 상 현

https://orcid.org/0000-0002-5196-6193
e-mail : sanghyun@yonsei.ac.kr
1989년 서울대학교 컴퓨터공학과(학사)
1991년 서울대학교 컴퓨터공학과 (공학석사)
2001년 UCLA 컴퓨터공학과(공학박사)
2001년~2002년 IBM T. J. Watson Research Center Post-Doctoral Fellow
2002년~2003년 포항공과대학교 컴퓨터공학과 조교수
2003년~2006년 연세대학교 컴퓨터과학과 조교수
2006년~2011년 연세대학교 컴퓨터과학과 부교수
2011년~현 재 연세대학교 컴퓨터과학과 교수
관심분야 : 데이터베이스, 데이터마이닝, 바이오인포매틱스, 빅데이터 마이닝 & 기계 학습