딥러닝 모델에서 증강 데이터를 활용한 선행 학습

김혁주○ 방채영 이동수 안홍렬*
수원대학교 데이터과학부
{khj4310_; co505123060; lds991016; hrahn}@suwon.ac.kr

Pre-training with Augmentation Data in Deep Learning Model

HyukJu Kim○ ChaeYoung Bang DongSu Lee Hongryul Ahn*
Division of Data Science, The University of Suwon

요 약

현대 사회에서 암은 사망 원인 1위에 해당하며, 암 치료제 개발을 위해서 기계학습 방법이 시도되고 있다. 본 연구에서는 유전자 발현 데이터에서 폐암의 아종을 분류하는 증강 데이터를 활용한 딥러닝 선행학습 기법을 개발하였다. 기존의 데이터를 조합하여 새로운 증강 데이터를 활용하는 것이 분류 모델의 정확도 향상의 핵심이다. 하지만 의료 분야에서는 데이터의 변형에 민감하기 때문에 증강 데이터와 같이 가공된 데이터를 학습에 직접 사용하는 것을 지양하는 것이 필요하다. 따라서, 증강 데이터로 모델을 선행 학습하고, 선행 학습된 모델 가중치 값을 본 학습 모델의 초깃값으로 설정하여 최종 모델에 대한 증강 데이터의 영향을 감소하였다. 개발한 방법을 폐암 아종 분류에 대한 유전자 발현 데이터에 적용하였을 때 더 높은 정확도를 보였다.

1. 서 론

2020년 발표된 국가암등록통계(국가승인통계 117044호)에 따르면, 암은 전체 사망 원인 중 1위이고, 특히, 폐암은 2018년 한 해 동안 우리나라에서 발생한 암 중의 11.7%(28,628건)를 차지하며, 전체 암 중에 2위에 해당하는 발생 빈도가 높은 암이다. 특히 폐암의 아종(subtype) 중 하나인 폐편평상피세포암(Lung Squamous Cell Carcinoma)은 기존 치료제 개발 방법으로는 개발의 어려움이 있어서 그 한 대안으로 기계학습 기법을 활용하여 암 환자 데이터를 분석하는 연구가 시도되고 있다.

그런데 기계학습이 효과적으로 이루어지기 위해서는 충분한 데이터가 필요하다. 하지만, 폐암 데이터는 샘플의 획득과 측정 비용 등의 문제로 데이터 개수가 1000개 내외로 보통의 다른 종류 데이터들보다 적기 때문에 기계학습 기법이 잘 적용되지 않는 문제가 있다. 본 연구에서는 이러한 적은 수 데이터인 유전자 발현 데이터에 대해서, 증강 데이터(Augmented data)를 이용하여 폐암 아종 분류에 대한 인공신경망 방법의 정확도를 향상하는 것을 목표로 한다.

2. 유전자 발현량 데이터에서의 폐암 아종 분류 문제

본 연구에서 2개의 폐암 아종(폐편평상피세포암, 폐선암)의 환자 1,018명에서 측정한 유전자 발현 데이터로 부터 폐암의 아종을 분류하는 기계학습 모델을 만드는 연구를 수행하였다. 데이터에서 폐암 환자는 폐선암 환자 517명(50.8%), 편평상피암 환자 501명(49.2%)으로 두 폐암 아종은 1% 차이 이내로 균등한 비율 분포를 이루고 있다.

유전자 발현 데이터를 기계학습을 위한 행렬 형태의 설명변수 데이터 와 반응변수데이터 Y 로 표현할 수 있다(그림 1). 설명 변수 데이터 X 는 N G 모양의 2차원 실수 행렬 데이터이다. 여기서, N 개의 행은 환자 샘플을 의미하고, G 개의 열은 유전자를 의미한다. 또, 반응 변수 데이터 Y 는 N 명의 환자에 대해서 각 환자가 폐선암(0)인지 폐편평상피세포암(1)인지를 나타내는 이진 값의 벡터이다. 즉, 아래 수식 또는 그림 1과 같이 정리할 수 있다.

X = {x_j : i번째 환자의 j번째 유전자에 대한 발현 실수값}

Y = <y_i : i번째 환자의 폐암 종류에 대한 이진값. 즉, 폐선암(0) 또는 폐편평상피세포암(1)>.

그림 1 폐암 환자의 유전자 발현 데이터

폐암 유전자 발현 데이터는 데이터 수가 적은 (N 은 1,000 내외) 소량 데이터로, 소량 데이터의 분류 문제에서 잘 동작하는 기계학습 방법의 개발이 필요하다.

3. 방 법

본 연구에서는 소량 샘플 유전자 발현 데이터에서 데이터 증강(Data augmentation)을 사용하여 데이터를 늘리고, 이 데이터를 2단계로 학습함으로써 분류 딥러닝 모델의 정확도를 향상하는 방법을 개발하였다.

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2020-0207).

3.1 데이터 증강 방법

데이터 증강 방법은 데이터의 수가 부족할 때, 기존 데이터를 변형 및 샘플링하여 새로운 데이터를 만드는 기법이다. 데이터 증강이 많이 활용되는 분야는 이미지 데이터에 대한 딥러닝 분석 분야로, 원본 이미지를 회전, 좌우 및 상하 반전, 밝기 조절 등 변환하여 증강 이미지 데이터를 생성(그림 2)하고 이 증강된 데이터를 모델 학습에 사용한다. 하지만 유전자 발현 데이터에서는 이미지 데이터에서의 회전, 반전, 밝기 조절 등의 변환이 정의되지 않기 때문에 이미지 데이터에서와는 다른 방식의 데이터 증강 방법이 필요하다.

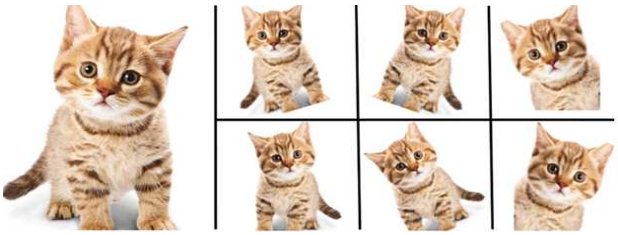

그림 2 원본 고양이 이미지 데이터(가장 왼 쪽)로 부터, 회전, 좌우 반전의 변환을 통해 생성한 증강 데이터(오른 쪽 6개)의 예시 [1]

본 연구에서는 수치형 유전자 발현 데이터 , Y에 대하여 통계적 방법(샘플링, 평균, 최빈값 등)을 활용하여 새로운 유전자 발현 데이터와 그 반응 변수 (, w)를 생성하는 세 가지 데이터 증강 방법을 제안한다.

그림 3 유전자 발현 데이터에서의 세 종류의 데이터 증강 방법

그림 3은 세 종류의 데이터 증강 방법을 보여준다. 첫 번째는 반응변수 값이 같은 샘플 n개를 랜덤 추출하여 샘플 집합 $S=\ s\ ,\cdots,s_n\ |\ y_1=\cdots=y_n\}$를 만들고, 샘플 집합 S에서 유전자의 값들을 평균하여 데이터를 생성한다. (즉, 유전자 j에 대해서, $v_j=\frac{1}{n}\quad s_{ij},\ w=y_1$).

두 번째는 반응변수 값에 상관없이 샘플 n개를 랜덤 추출하여 샘플 집합 $S=\{s_1,\cdots,s_n\}$를 만들고, 샘플 집합 S에서 유전자의 값들은 평균, 반응변수 값은 최빈값으로 데이터를 생성한다. (즉, 유전자 j에 대해서, $v_j=\frac{1}{n}\sum s_{ij}$, $w=\mathrm{mode}(y)$).

세 번째는 반응변수 값에 상관없이 샘플 n개를 랜덤추출하여 샘플 집합 $S=\{s_1,\cdots,s_n\}$를 만들고, 샘플 집합 S에서 유전자의 값들을 평균하여 데이터를 생성한다. (즉, 유전자 j에 대해서, $v_j=\frac{1}{n}\sum s_{ij},\ w=\frac{1}{n}\sum yw=y_1$).

3.2 증강 데이터를 사용한 딥러닝 모델 사전 학습 (pre-training)

사람의 시각 인지 능력을 통해 품질을 판별할 수 있는 이미지 데이터의 경우, 바르게 생성된 증강 이미지 데이터만 선별하여 모델 학습에 사용할 수 있다. 하지만, 비지각적 데이터인 유전자 발현 데이터는, 현재 기술로는 증강 데이터가 바르게 생성되었는지 판단하는 것이 불가능하다. 따라서 생명 윤리와 관련된 의료 데이터 분석에 증강 유전자 발현 데이터를 사용할 때에는 생성된 증강 데이터가 최종 모델에 영향이 적게 가지도록 하는 신중하게 적용하는 것이 필요하다.

우리는 증강 데이터를 기계학습 모델 학습에 활용하면서도 최종 모델에 영향력은 최소화하는 2단계 기계학습 방법을 고안하였다(그림 4). 딥러닝 학습 과정은 다음 두 단계로 나누어진다. 1단계 학습에서는 최초에 딥러닝 모델의 가중치를 랜덤하게 초기화한 이후에 증강 데이터와 원본 데이터를 합친 데이터를 사용하여 초벌 학습하고 2단계 학습에는 원본 데이터만 사용하여 최종 학습한다. 이러한 2단계 학습 기법은 증강 데이터를 활용하여 더 많은 데이터로 학습하는 이점을 얻으면서도 증강 데이터의 최종 모델에서 영향력을 감소시키고 원본 데이터의 최종 모델에서 영향력을 증가시키는 효과를 가진다.

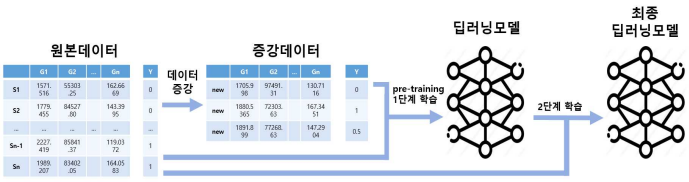

그림 4 증강 데이터를 활용한 2단계 딥러닝 모델 학습 기법

3.3 딥러닝 모델

이 논문의 실험에서 사용한 딥러닝 모델의 구조와 하이퍼 파라미터는 다음과 같다. 본 연구에서는 딥러닝 모델 구조로 100개의 은닉 유닛을 가지는 1개의 은닉층을 가지는 multi-layer perceptron을 사용하였다. 목적 함수로는 cross entropy와 alpha 0.1의 L2-regularization의 합을 사용하였고, 가중치 초

기화 방법은 $\left(\frac{1}{n_n}, \frac{1}{n_{in}}\right)$를, 최적화 알고리즘으로는 learning rate 0.001의 adam optimizer를 사용하였다.

4. 실험 및 결과

4.1 데이터 처리

본 연구에서 사용된 데이터는 미국의 TCGA 프로젝트[2]의 폐암 환자의 유전자 발현 데이터로서, fireblowser 데이터베이스(http://firebrowse.org/)를 통해서 다운로드 받았다. 우리는 암 관련 데이터베이스인 MsigDB[3]를 참고하여 전체 19,977개의 유전자 중에서 암과 관련된 323개의 유전자를 선택하여 데이터 차원을 축소하였다. 그 후 각 유전자에 대해서 log2 변환과 z-score 변환을 수행하여 전처리하였다. 본 실험에서 인공신경망 모델은 python 언어의 pytorch 라이브러리를 사용해서 프로그래밍 되었으며, 성능 비교의 알고리즘은 python scikit-learn, XGboost[4], lightGBM[5] 라이브러리의 코드를 사용하였다. 실험은 google colab 환경에서 수행되었다.

4.2 실험 방법

본 실험에서는 scikit learn 라이브러리의 StratifiedKFold 함수를 사용하여 10-fold 교차검증 실험을 수행한다. 각 교차 검증에서 유전자 발현 데이터를 통해 폐암 아종을 분류하고 정확도(accuracy)를 측정한다. 비교 방법으로는 sckitlearn 라이브러리에 구현된 베이스라인 MLP, Logistic Regression[6], AdaBoost[7], Bagging[8], Random Forest[9], SVM[10], XGboost, LightGBM의 알고리즘과 비교되었다.

4.3 실험 결과

먼저 우리는 고안한 세 가지 데이터 증강 방법에 대해서, 세 종류의 데이터 증강 방법, 샘플링 개수 n, 증강 데이터 배수 r의 파라미터를 변화시키면서, 파라미터에 대한 정확도 변화를 테스트하였으며, 두 번째 증강 방법에서 4개의 샘플을 추출하여 train set의 크기만큼 증강 데이터를 생성하는 방법의 성능이 가장 우수하다는 것을 확인하였다.

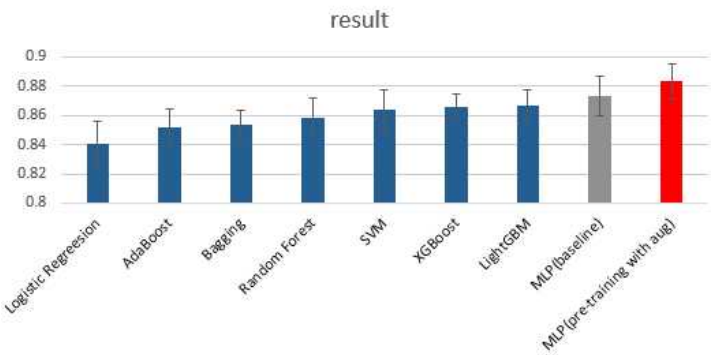

그림 5 폐암 분류 문제에 대한 모델의 정확도 비교. 에러 막대는 표준 오차(standard error of the mean)이다.

우리는 폐암 아종 분류에 대한 정확도에 대해 개발한 알고리즘과 8개의 기계학습 모델 알고리즘을 비교하였다(그림 5). 증강 데이터로 선행 학습한 딥러닝 MLP 모델은 88.32%로 가장 높은 정확도를 보였다. 특히, 2등인 베이스라인 MLP 모델의 정확도인 약 87.33%보다 약 1% 정도 상승함으로 증강 데이터를 사용한 사전학습이 효과가 있음을 확인할 수 있었다.

5. 결론

본 논문에서는 유전자 발현 데이터를 증강하고 2단계 학습을 수행하여 폐암 아종 분류 딥러닝 모델의 분류 정확도를 향상하는 방법을 제시하였다. 증강 데이터를 사전학습에서 사용하고 원본 데이터를 최종 학습에서 사용하는 우리의 방법은 최종 모델의 폐암 아종 분류 문제에서의 정확도 향상에 유효한 효과를 나타내었다.

참고문헌(Reference)

[1] https://insighting.tistory.com/13

[2] Weinstein, John N., et al. "The cancer genome atlas pan-cancer analysis project." *Nature genetics* 45.10 (2013): 1113-1120.

[3] Liberzon, Arthur, et al. "Molecular signatures database (MSigDB) 3.0." *Bioinformatics* 27.12 (2011): 1739-1740.

[4] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016.

[5] Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." *Advances in neural information processing systems* 30 (2017): 3146-3154.

[6] Zhu, Ciyou, et al. "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization." *ACM Transactions on Mathematical Software (TOMS)* 23.4 (1997): 550-560.

[7] Freund, Yoav, and Robert E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting." *Journal of computer and system sciences* 55.1 (1997): 119-139.

[8] Breiman, Leo. "Bagging predictors." *Machine learning* 24.2 (1996): 123-140.

[9] Breiman, Leo. "Random forests." *Machine learning* 45.1 (2001): 5-32.

[10] Platt, John. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods." *Advances in large margin classifiers* 10.3 (1999): 61-74.