비순차 데이터 입력에 대한 Convolution Neural Network 분류기 적합성 평가

왕창원*, 석현석**, 신항식*,***
*서울아산병원 의공학연구소, **전남대학교 바이오메디컬공학 협동과정, ***울산대학교 의과대학 서울아산병원

Evaluation of Convolution Neural Network classifier compatibility for Non-Sequential Data Input

Changwon Wang*, Hyun Seok Seok**, Hangsik Shin***
*Biomedical Engineering Research Center, Asan Medical Center, Seoul, Korea
**Dept. of Biomedical Engineering, Chonnam National University, Yeosu, Korea
***Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea

Abstract - CNN(Convolution Neural Network)은 입력 데이터의 순차적 입력 특성 및 연결성을 잘 반영하기 때문에, 이미지나 영상 데이터의 분류연구에 널리 활용되고 있다. 본 연구에서는 비순차적 특성을 가지는 입력에 대한 CNN의 적합성을 검증하고자 한다. 독립적인 feature를 갖는 입력 데이터의 순차적 입력 특성 및 연결성을 없애기 위해 무작위 순열(permutation)을 수행한 뒤, 1D CNN 및 RandomForest에 입력 데이터로 사용하여 각 모델의 적합성을 비교 검증하였다. 분석 결과, 1D CNN에서는 AUC 0.77 ± 0.05, RandomForest에서는 0.94 ± 0.00로 나타났다. 1D CNN은 무작위 순열된 입력 데이터에 대한 표준편차가 RandomForest보다 높게 나타났으며, 입력 데이터의 순서에 따라 성능 편차가 큰 것을 확인하였다. 본 연구의 결과는, 비순차적 특성을 갖는 입력 데이터를 CNN에 적용하는 연구의 모델 적합성 평가 및 최적의 모델을 도출하는 데 있어 도움이 될 것으로 기대된다.

1. 서 론

기존 비순차적 입력 데이터에는 SVM(Support vector machine), RandomForest 모델 등을 사용하였으나 [1], 최근 비순차적 입력 데이터를 CNN에 적용하는 연구가 시도되고 있다 [2].

CNN은 콘볼루션 과정을 통해 입력 데이터의 순차적 입력 특성과 연결성을 잘 반영하기 때문에 이미지나 영상 같은 데이터 분류 관련 연구에 많이 활용되고 있으며, 높은 분류 성능을 갖고 있는 것으로 보고되고 있다 [3-4].

하지만, CNN은 콘볼루션 과정 중 kernel size에 따라 입력 데이터의 순차적 특성과 연결성이 잘 반영되기 때문에, 비순차적 특성을 갖는 입력 데이터를 사용한다고 가정했을 때 입력 데이터 순서에 따라 성능의 편차를 보일 수 있다.

따라서, 본 연구에서는 비순차 입력 데이터에 대한 CNN 적합성을 평가하기 위해, 무작위 순열을 통한 비순차적 특성과 연결성을 없앤 입력 데이터를 CNN과 RandomForest에 적용하여 각 모델의 적합성을 검증하였다.

2. 본 론

2.1 데이터 세트

데이터세트는 중국 Wuhan의 Union, Liyuan 병원에서 수집한 HUST-19 open database를 사용하였다 [5]. HUST-19는 Covid-19 환자 및 대조군 1,521명의 인구학적 정보와 Routine Blood Test, Inflammation, Biochemical Test, Blood coagulation Test, Immune cell Typing, Cytokine Profile Test에 대한 정보가 포함되어 있으며, 총 86개의 feature로 구성되어 있다. 본 연구에서는, 대조군 및 그룹이 미분류된 피험자를 제외한 719명 중 사망자 그룹 피험자 57명이었으며, 데이터 세트 비율을 1:1로 맞추기 위해 SARS-Cov-2 nucleic acids가 positive를 보인 57명을 선정하였다. 최종적으로 114명(치료된 그룹 : 57명, 사망한 그룹 : 57명)의 데이터를 사용하였다. 그룹 간, 연령에 대한 차이가 있는지 확인하기 위해 카이제곱 검정을 수행하였으며, 검정 결과 연령분포에 대한 차이는 없는 것을 확인하였다 (표 1).

<표 1> 그룹 간 Age feature에 대한 카이제곱 검정

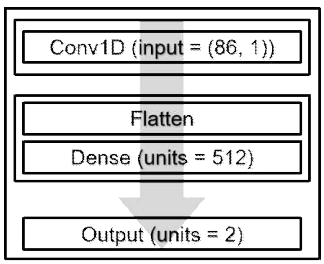
	χ^2	dof	p-value
Age	71.24	56	0.08

Training 및 test 데이터는 각각 79, 35명으로 구성하였으며, Training 데이터 중, 30%인 24명은 validation 데이터로 사용하였다. Mortality prediction을 위해 치료된 그룹과 사망한 그룹으로 labeling 하였다.

2.2 CNN 및 Random Forest 모델

본 연구에서 사용한 CNN 및 RandomForest 모델은 그림 2와 같이 구성하였다. 무작위 입력 데이터에 대한 CNN의 적합성을 검증하기 위해 1층 1D CNN 모델은 콘볼루션 계층, 완전 연결계층, 결과 출력 계층 순으로 구성하였고, 콘볼루션 필터 개수는 32개, 완전 연결계층의 입력 크기는 512, batch 크기는 10, epoch 크기는 300으로 설정하였다.

RandomForest 모델에서 트리는 10개, 최대 트리 깊이(max tree depth)는 2부터 20까지 2씩 증가시키면서 총 10개의 최대 트리 깊이를 사용하여 모델을 구성하였다.



<1D CNN Model>

<그림 1> 1D CNN 모델

2.3 데이터 분석

데이터 분석은 총 3가지로 나누어 수행하였으며, 무작위 순열 입력을 100번 반복하면서 얻은 결과를 분석에 사용하였다.

첫 번째 분석에서는 1층 1D CNN 및 RandomForest 간 무작위 순열 검정을 수행하였다. 두 번째 분석에서는, 1층 1D CNN의 kernel size를 변경해가며 AUC 평균 및 표준편차 분석을 수행하였다. 이때, kernel size는 전체 feature 길이의 10%부터 100%까지 10%씩 증가시켜가며 얻은 kernel size를 사용하였다. 마지막 분석에서는, RandomForest의 트리 깊이에 따른 AUC 평균 및 표준편차를 분석하였다.

3. 결　　과

3.1 1D CNN 및 RandomForest의 무작위 순열 검정

무작위 순열 검정 결과를 보면, 1D CNN에서는 AUC 평균 및 표준편차는 0.78 ± 0.05로 나타났으며, RandomForest에서는 0.94 ± 0.02로 나타났다 (표 2).

<표 2> 무작위 순열 유무에 따른 AUC 평균 및 표준편차

	CNN	RandomForest
mean AUC(SD)	0.78 (0.05)	0.94 (0.02)

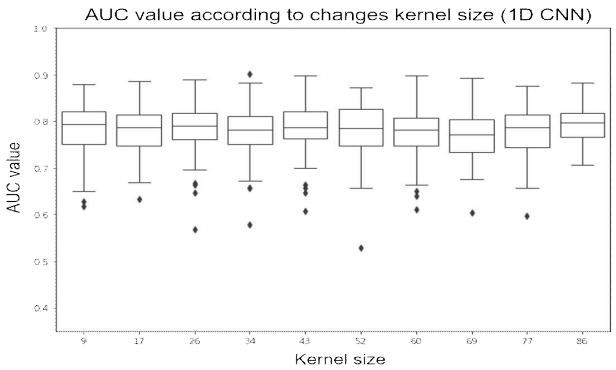
3.2 Kernel size 별 AUC 평균 및 표준편차

1D CNN의 Kernel size 별 AUC 평균 및 표준편차 결과를 보면, kernel size가 가장 작은 9인 경우 AUC 평균 및 표준편차는 0.79 ± 0.06으로 나타났고, 가장 큰 86인 경우 0.80 ± 0.04로 나타났다 (그림 2, 표 3).

<표 3> Kernel size 별 AUC 평균 및 표준편차

	Kernel size according to the ratio of the number of features (%)									
	10	20	30	40	50	60	07	80	09	100
mean AUC (SD)	0.79 (0.06)	0.78 (0.05)	0.79 (0.05)	0.78 (0.05)	0.79 (0.05)	0.78 (0.05)	0.78 (0.05)	0.77 (0.05)	0.78 (0.05)	0.80 (0.04)

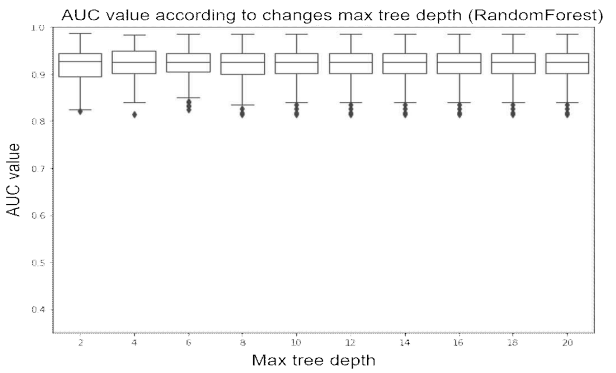
**AUC: area under curve, SD: standard deviation*



<그림 2> 1D CNN의 Kernel size 별 AUC 및 표준편차

3.3 최대 트리 깊이 변화에 따른 AUC 평균 및 표준편차

그림 3, 표 4는 최대 트리 깊이의 변화에 따른 RandomForest의 성능을 나타낸 것으로, 트리의 최대 깊이별 AUC 평균은 0.92로 동일하게 나타났다. 표준 편차는 트리의 깊이가 6일 때 0.03으로 가장 적게 나타났으며, 나머지에서는 0.04로 동일하게 나타났다.



<그림 3> RandomForest의 최대 트리 깊이 별 AUC 평균 및 표준 편차

<표 4> 최대 트리 깊이 변화에 따른 AUC 평균 및 표준편차

	Max tree depth of Random Forest Classifier									
	2	4	6	8	10	12	14	16	18	20
mean AUC (SD)	0.92 (0.04)	0.92 (0.04)	0.92 (0.03)	0.92 (0.04)	0.92 (0.04)	0.92 (0.04)	0.92 (0.04)	0.92 (0.04)	0.92 (0.04)	0.92 (0.04)

**AUC: area under curve, SD: standard deviation*

4. 결　　론

본 연구에서는 비순차적 입력 데이터에 대한 CNN의 적합성을 검증하기 위해, 1층 1D CNN 및 RandomForest에 무작위 순열된 입력 데이터를 사용하고 AUC 평균 및 표준편차를 비교분석 하였다.

두 모델의 무작위 순열 검정 결과를 보면, 1D CNN은 RandomForest보다 AUC 평균이 낮고, 표준편차가 더 크게 나타났으며 입력 데이터의 순차적 연결성에 따라 성능 편차가 큰 것으로 나타났다.

1D CNN의 kernel size가 작을수록 AUC 분포의 편차가 크게 나타났으며(SD=0.06), 커질수록 분포는 작게 나타나는 것을 확인하였다(SD=0.04). 즉, 무작위 순열된 비순차 입력 데이터를 CNN에 적용할 경우, 입력 데이터의 순서 및 연결성의 변화에 따라 성능 편차가 클 수 있음을 확인하였다.

RandomForest의 최대 트리 깊이 변화에 따른 AUC 평균 및 표준편차를 보면, 트리의 깊이 변화에 따른 AUC 평균은 동일하게 나타났으며, 표준편차 또한 트리의 값이 6인 경우를 제외하고 일정한 성능을 보이는 것으로 나타났다.

본 연구 결과를 통해, 비순차적 입력 데이터에 대해서는 CNN보다 RandomForest에서 더 안정적이며 높은 AUC 및 적은 편차를 갖는 것으로 나타났다.

향후, 두 모델의 보다 정확한 비교를 위해, 모델의 하이퍼 파라미터 최적화 및 다양한 케이스에 대한 추가적인 확인이 필요할 것으로 생각된다.

본 연구의 결과는, 비순차적 입력 데이터를 1D CNN에 적용할 때 모델의 적합성 평가 및 최적의 성능을 갖는 모델을 도출하는데 있어 도움이 될 것으로 기대된다.

감사의 글

본 연구는 2021년도 보건복지부 재원으로 질병중심중개연구사업(HI21C0011) 및 한국보건산업진흥원의 보건의료기술 연구개발사업(HI18C2383)의 지원을 받아 수행하였습니다.

[참 고 문 헌]

[1] RC. Chen, C. Dewi, SW. Huang et al. “Selecting critical features for data classification based on machine learning methods,” *J Big Data,* Vol. 7, p. 52, 2020.

[2] C. Potes, S. Parvaneh, A. Rahman, B. Conroy “Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds,” *IEEE 2016 Computing in Cardiology Conference*, Vol. 43, p. 621-624, 2016.

[3] L. Alzubaidi, J. Zhang, A. J. Humaidi et al. “Review of deep learning: concepts, CNN architectures, challenges, applications, future directions,” *J Big Data*, Vol. 8, p. 53, 2021.

[4] Y. Shin, I. Balasingham, “Comparison of hand-craft feature based SVM and CNN based deep learning framework for automatic polyp classification,” *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, p. 3277-3280, 2017.

[5] W. Ning, S. Lei, J. Yang et al. “Open resource of clinical data from patients with pneumonia for the prediction of COVID-19 outcomes via deep learning,” *Nat Biomed Eng*, Vol. 4, p. 1197 - 1207, 2020.