Deep Learning Final Project

-신경망을 사용한 심전도데이터 LVH 분류모델-

3조

2016311924 양지연

2016314177 강유성

2016312623 이성준

1. 문제정의

1) 모델의 목적

우리의 궁극적인 목적은 심전도 데이터를 입력받아 좌심실 비대의 유무를 높은 성능으로 판별하는 모델의 구축하는 것이다.

2) 왜 좌심실 비대인가?

좌심실 비대증(LVH)은 초기에는 증상을 느끼지 못하지만, 심혈관 질환 발생 예측의 중요한 전조 질환으로, 심전도 검사의 주된 목적 중 하나이다. 2005년 가정의학회지에 기재된 '좌심실 비대의 진단 방법으로서 심전도'에 따르면 LVH의 경우, 심전도검사에서 민감도는 낮고 특이도는 높은 경향을 보인다. 즉, 심전도검사를 통하여 LVH로 의심될 확률이 낮다는 것이다. 질병의 진단 경우, 실제로는 병을 앓고 있지만 그렇지 않다고 진단하는 것은 큰 문제이다. LVH를 더 정확하게 진단하는 방법으로는 심장 초음파 검사가 있지만, 비용이 많이 들고 일차의료에서는 누구나 이용할 수 없어 보편화하기 어렵다는 단점이 있다. 그렇기 때문에 일차 의료에서 행해질 수 있는 간단한 1차 검사인 심전도 검사 데이터에서 LVH를 잘 분류해내는 것은 빠르게 심장에 이상이 있음을 감지하고 한 발 더 빠른 의학적 조치를 가능하게 한다.

또한 우리가 보유한 데이터에서 LVH 유무에 대한 분포를 확인해 보면 108대 92로, 두 개의 클래스가 상당히 균형 잡힌 데이터 분포를 보인다. 데이터가 불균형할 경우, 하나의 클래스로 편향된 분류모델이 될 위험이 크다. 이를 통해 한정된 데이터를 통해 학습시킬 수 있는 분류모델 중 LVH를 분류하는 모델이 이런 위험을 방지할 수 있는 좋은 모델이라고 판단하였다.

3) Input Data 정의

Anno와 anno_idx는 signal 리스트 데이터 중 각 파형의 위치에 해당하는 인덱스에 대한 정보를 가지고 있는 데이터로, 5,000의 차원을 가지고 있는 시그널데이터에서 핵심적인 정보를 뽑아낼 수 있는 잠재적 변수라고 판단하였다. 대부분의 anno는 (n)(t)(p)의 순서로 반복되는 양상을 띠고 있는데, 이에 따라 반복되는 (n)(t)(p)의 시그널값들을 평균한 값을 한 record의 시그널 추세를 대변하는 값으로 활용하기로 하였다. 결론적으로 age, sex 변수는 categorical 변수로 취급하여 one-hot encoding 한 변수들로 변환하고, 각 파형의 시작, 피크, 끝의 평균적인 값들을 저장한 9개의 element로 이루어진 list를 저장해주어 유도 별 9개의 리스트로 이루어진 signal_ntp 변수를 추가해 총 117개의 차원으로 input을 결정하였다. 하지만 심전도 데이터는 주기적인 파동을 지니고 있는 신호 데이터이기에 평균적인 값들로 나타낸 변수로는 놓치는 부분이 있으리라 판단하여 5,000개의 시그널값을 전부 사용하는 또 다른 입력값도 사용할 예정이다. 시그널 값을 전부 사용하는 입력값의 경우, 12개의 유도 값을 전부 사용하는 것이 아니라 LVH의 진단기준인 Sokolow-Lyon index를 참고하여 V1, V5, V6, aVL 4가지 유도만을 사용하여 학습의 효율성을 늘리고자 하였다.

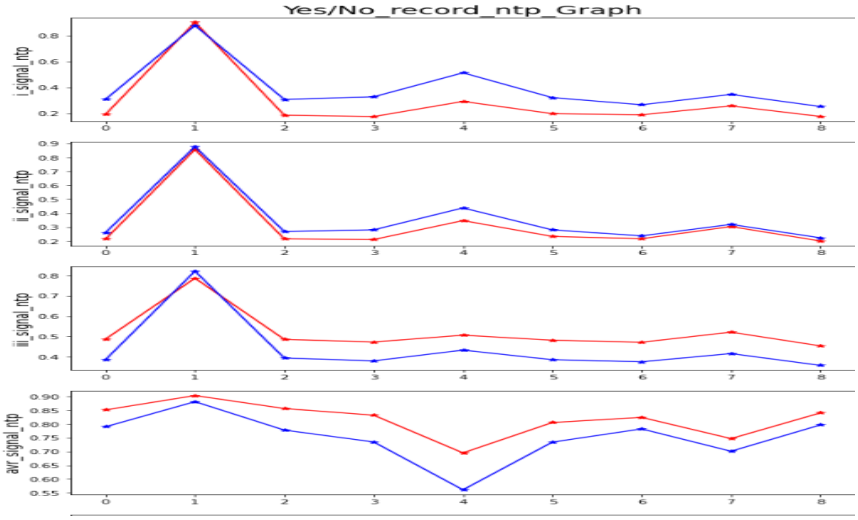
4) 평가지표

모델의 대표적인 평가지표인 정확도는 불균형한 데이터의 경우, 즉 하나의 클래스가 희박할 경우 모든 예측을 다수의 클래스로 예측한다 해도 높은 정확도를 출력한다. 하지만 LVH의 경우 데이터가 비교적 균형 잡혀 있기 때문에, 우리가 학습할 모델의 성능지표로 활용할 것이다. 하지만 데이터 자체의 수가 너무 작기 때문에 좀 더 정확한 성능평가를 위하여 precision과 recall, 그리고 이 두 가지를 모두 고려한 average precision 또한 성능평가지표로 사용할 예정이다. Precision과 recall은 서로 반비례하기 때문에 이 두 가지를 모두 고려할 수 있는 AP가 필수적이며, 실제로 물체 검출 및 이미지 분류 알고리즘의 성능지표로 많이 사용되고 있다.

2. 팀에서 정한 모델을 사용하는 이유

1) DNN BInary-Classifier

밑의 그래프는 input data에서 언급하였던 파형의 특징을 함축한 record_ntp를 LVH 유무에 따라 시각화한 것 중 일부이다. 그래프를 보면 LVH유무에 따라 파형에 차이가 있음을 알 수 있다. 이를 통해 5,000개의 시그널을 대신할 새 변수가 LVH 여부를 분류하는데 충분한 학습자료가 될 수 있음을 확인하였다. 이러한 input 데이터를 바탕으로 입력변수 간의 비선형조합이 가능하며 예측력이 다른 머신러닝 기법들보다 비해 상대적으로 우수한 DNN BInary-Classifier를 모델링 하였다. 3개의 hidden layer를 쌓았으며 hidden layer의 활성 함수로는 +/-가 반복되는 교류에서 –의 흐름을 차단하는 'ReLu'를 사용하였다. 마지막 출력층에는 이진 분류를 위한 'Sigmoid' 함수가 사용되었다. 또한 추가적인 평가 기준 설정을 위한 compile 함수에는 이항 간 분류 문제를 다루는 'BinaryCrossEntropy'와, 지그재그 현상이 줄어들고 관성의 효과를 낼 수 있는 momentum의 장점과 최근 값과 이전 값에 각각 가중치를 주어 최근 값을 더 잘 반영할 수 있는 RMSprop의 장점이 적절히 조화된 'Adam optimizer'를 사용하였다. 마지막으로 층을 지날수록 weight가 너무 커지거나 작아지는 것을 방지하기 위해 hidden layer사이에 'Batch Normalization'을 사용하였다.



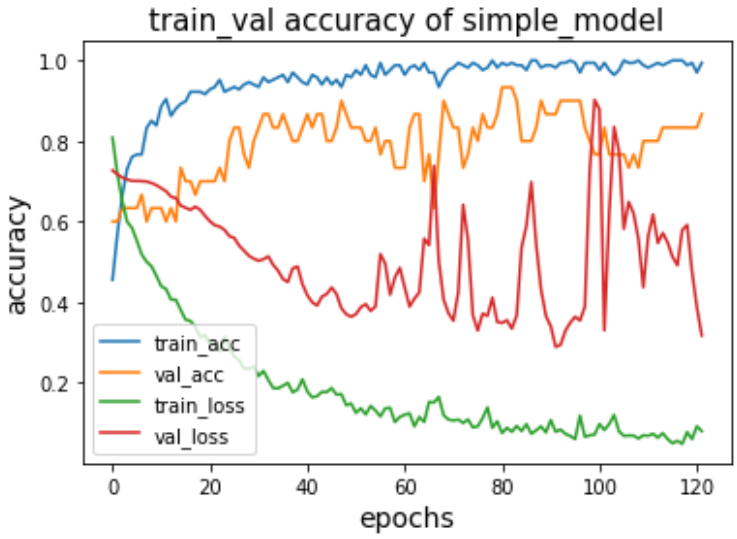
2) 1-D CNN Binary-classifier

앞에서 언급했듯이 핵심적인 특성만 추출한 새로운 변수는 학습에 있어서 놓치는 부분이 있을 수 있다. 이런 상황을 보완하기 위해 4개의 유도의 시그널값을 학습하는 새로운 모델이 필요하였다. 많은 수의 데이터를 효과적으로 학습하기 위해 우리는 1-D CNN을 채택하였다. CNN은 각 레이어의 입출력 데이터의 형상을 유지하고, 필터를 공유 파라미터로 사용하기 때문에 일반 신경망과 비교하여 학습 파라미터가 매우 적다. 또한 많은 연구들이 심전도 데이터 분석을 위해 CNN 모델을 사용하는 것을 확인하였고 그 중 2004년 IOP Conf.Series: Journal of Physics에 기재된 'Deep Learning for ECG Classification'을 참고하여 1D-CNN을 선정하였다. 1D-CNN은 고정적인 길이의 특징을 뽑아내는데 특화되어 있으며, 위치에 덜 민감한 정보를 잘 처리한다. 우리가 학습할 데이터의 경우 시간에 따라 5,000개의 심전도 전압으로 고정되어 있으며, 전압값이 중요하므로 위치에 덜 민감한 1D-CNN이 적당하다고 판단하였다. 총 5개의 hidden layer를 쌓았으며, CNN 함수의 커널 사이즈는 (250, 4), strides는 (10, 1)로 설정하였다. 그 이유는 보통 사람의 심박 수가 1분에 60~120 사이에 위치한다는 것과 초당 500개의 전압을 측정한다는 것을 고려하여, 놓치는 정보를 최대한 줄이기 위해서이다. 또한 우리는 차원의 수가 줄어들기를 원하기 때문에 valid Padding으로 데이터 차원의 축소를 막는 padding을 사용하지 않았으며, 위의 DNN과 같은 이유로 'Batch normalization'과 'Relu'함수를 이용해 출력값을 조절하였다. 추가로 시그널은 10초 동안 반복되기 때문에 중복된 특징이 많아 층마다 0.5의 비율로 'drop out'을 사용하여 입력 벡터를 줄였다. 이 CNN 또한 최종적으로 이진 분류모델이기에 출력층의 활성 함수와 model.compile함수의 파라미터들은 DNN과 똑같이 구성하였다.

3. 모델의 성능을 높이기 위해 시도한 내용

1) Early stopping

DNN 모델의 경우 train accuracy와 loss는 계속해서 개선되는 데 비해, validation accuracy와 loss는 가면 갈수록 오히려 개선되지 않는 양상을 보인다. 이는 train set에 모델이 overfitting 되고 있다는 증거로 이를 막기 위해, train accuracy와 validation accuracy의 추이를 고려하여 epoch를 끝까지 돌지 않고 과적합 되지 않는 적정한 선에서 모델을 먼저 멈춰주는 early stopping callback 함수를 활용하였다. 아래의 그래프는 early stopping 함수를 사용하여 학습한 결과값을 시각화한 것이다. 그래프를 보면 어느 정도 train accuracy가 지속적으로 높아지는 지점에서, validation accuracy와 큰 성능의 차이를 보이지 않는 지점에서 알아서 끊어주고 있음을 볼 수 있다. 200개의 epoch을 돌 수 있도록 설정하였는데, early stopping 함수로 인해, 적정선인 122번째의 epoch에서 먼저 멈추었으며, 결론적으로 train accuracy는 0.994의 성능을, val_accuracy는 0.8667로 더 개선된 결과를 낼 수 있었다.



CNN 모델의 경우, 똑같은 양상을 보였으며 DNN에 비해 빠르게 과적합 되는 것을 확인하였다. Early stopping 함수를 통해 성능의 개선은 이루어지지 않았지만, 빠르게 epoch을 멈춰 시간 효율성을 높일 수 있었다.

2) K-fold Cross Validation

별도의 test set이 없기 때문에 현재 만들어진 모델의 성능이 나쁘지 않다고 판단되어도, 실제로 일반화시킬 수 있을지의 여부는 확신할 수 없다. 결과적으로는 validation set에도 오버피팅되었을 가능성이 있기 때문이다. 또한 일반적으로 train validation 세트를 나누었을 때 train 세트에 분류하기 어려운 샘플들만 담기는 경우, 혹은 test 세트에 분류하기 쉬운 샘플들만 담기는 경우처럼, 평가 지표를 측정할 때 편파적으로 세트들이 나뉘는 경우가 생길 수 있기 때문에 전반적인 데이터들을 모두 고려하여 모델을 구성하였다고 이야기하기 어렵다. 이러한 이유로 일반화의 성능을 측정하기 위해 하나의 rain/validation 세트를 나누는 것보다 더 안정적인 방법인 K-fold Cross Validation 방법을 사용하여 모델 학습을 진행하였다. K-fold Cross Validation을 통하여 데이터를 7개의 fold로 나누어 모든 데이터를 train과 test set에 한 번씩 활용되는 것을 기대할 수 있으며, 이를 통해 데이터의 수가 너무 작다는 단점을 보완하고 성능의 향상을 기대할 수 있다. 실제로 CNN의 경우 기존 accuracy가 0.825인 train/test set의 결과값이 0.94로 상당히 개선되었음을 확인할 수 있었다.

3) CNN 모델의 파라미터 조절

마지막으로 현재 모델보다 성능을 좀 더 개선하기 위해 커널 사이즈와 stride의 폭을 조절하였다. 우선 커널 사이즈를 조절한 후 결과값을 비교해 최선의 값을 선정한 후 stride의 폭을 조절하였다. 커널 사이즈는 기존의 250을 두 배 값인 500과 그 사이 값인 400을 가지고 비교해 보았다. 그 결과

500은 기존보다 성능이 더 저조했으며 400은 기존 값과 비슷한 성능을 보여 기존의 250을 선정하였다. Stride의 폭은 커널의 크기를 넘지 않아야 하므로, 250 안에 있는 5의 배수들인 10, 50, 125를 가지고 비교해보았다. 그 결과 stride의 폭이 10인 경우 accuracy 0.95, AP 0.94로 성능이 기존보다 개선되었으며, 1/25씩 움직이면서 특징을 잡았을 경우 가장 특징을 잘 추출한다는 것을 알 수 있었다.

4. 결과의 분석

1) DNN과 CNN의 최종결과 비교

핵심적인 특징을 뽑아 학습시킨 DNN 모델의 경우 최종적으로 79.6%의 정확도를 보였으며, 4개의 핵심적인 유도들의 모든 시그널을 학습시킨 CNN의 경우 최종적으로 95.51%의 정확도와 94.4%의 AP값을 얻을 수 있었다. DNN의 경우 fold에 따른 모델의 성능 차가 나는 것으로 보아, 모델이 train, validation set의 형태, 특성에 따라 영향을 많이 받는 것을 알 수 있다. 또한, 최종적인 성능도 그렇게 높지 않아 5,000개의 시그널을 9개로 축소하는 과정에서 데이터 손실이 발생해 모델 성능의 저하로 이어진 것으로 예상된다. 또한 시그널 데이터가 특정한 주기에 따라 반복되는 시계열성 데이터임에도 불구하고, 해당 DNN 모델은 이러한 주기적인 파동 정보들을 고려하지 못하기 때문에 추가적인 정보 손실이 생겨난 것으로 예상된다. 이와 비교해 5,000개의 시그널값을 입력으로 학습한 CNN은 5,000개의 시그널 데이터의 특징을 잘 반영하여 차원을 축소해 학습하여 앞에서 언급한 DNN 모델의 단점을 극복하여 좋은 성능을 보여주었다. 결과적으로 커널 사이즈와 stride 폭의 파라미터 조절로 최고의 성능을 보인 1-D CNN 모델이 최종 모델로 선정되었다.

2) 의의 및 한계점

심전도 데이터를 사용하여 좌심실 비대를 분류하는 최종 모델을 선정하는 데 있어서, 데이터를 다방면으로 활용해 특히 전처리와 시각화에 큰 노력을 기울였다. 또한 다양한 input data 형태를 활용하여 걸맞은 모델에 각각 적용하였고 K-fold CV를 사용하는 등 적은 양의 데이터를 활용하여 과적합 문제를 고려하기 위해 노력하였다. 그 결과 높은 성능을 가진 최종 모델을 선정할 수 있었다.

하지만 DNN 모델의 경우, LVH 진단에 결정적인 기준이 되어주는 절대적인 값을 반영해주는 변수를 추가했다면 더 나은 성능을 기대할 수 있었을 텐데 데이터를 normalize 하는 과정 때문에 적용하기가 어려웠다. 그리고 CNN 모델은 파라미터 조정에서 더 많은 경우의 수를 고려하지 못한 것이 한계점이다. 또한 데이터가 너무 작았다는 점, 다양한 input 데이터 형태의 경우의 수를 고려하지 못한 점이 상당히 아쉬웠다.