Motor imagery EEG 분류

서울과학기술대학교 산업정보시스템전공

22101911 왕현식, 데이터 전처리 및 분석, 보고서 작성

22101942 한다은, 데이터 전처리 및 분석, 보고서 작성

요약

본 프로젝트의 목적은 Kaggle에서 제공된 EEG 데이터를 사용하여 4가지 종류의 motor imagery를 분류하는 머신 러닝 모델을 개발하고 성능을 평가하는 것이다. 사용된 데이터는 train과 test 데이터로 나뉘어 제공되었으며, 각각의 데이터는 22개의 전극에서 측정된 4608개의 샘플을 포함한다.

분석 과정에서는 라벨별로 푸리에 변환 시각화를 통해 데이터 차이를 분석했다. 특정 채널이 bad channel로 식별되어 제거되었으며, 주요 대역의 밴드 파워를 계산하여 시각화했다. 델타와 알파 대역이 라벨 별 차이를 보였고, 이를 사용하여 모델링을 진행했다.

CNN 모델은 3개의 컨볼루션 레이어와 풀링 레이어, 3개의 fc레이어 및 드롭아웃 레이어로 구성되었다. 모델 훈련 결과, train loss와 valid loss, valid accuracy를 비교했으며, 최종적으로 test 데이터에 적용한 결과 캐글 제출 점수는 0.2934였다.

모델 개선을 위해 CWT와 PSD를 사용하여 신호를 변환하고 새로운 CNN 모델을 생성했다. 이 모델은 50 에폭 동안 학습되었으며, valid accuracy 0.4154를 달성했다. 개선된 모델을 test 데이터에 적용한 결과 캐글 제출 점수는 39.583%로 향상되었다. 하지만 여전히 높은 성능을 보이지 못하는 한계가 있어, 추가적인 전처리의 중요성을 시사한다.

1. 서론

본 프로젝트의 목적은 kaggle에서 제공된 EEG 데이터를 활용하여 4가지 종류의 motor imagery를 분류하는 머신 러닝 모델을 개발하고 성능을 평가하는 것이다.

2. 사용 데이터

kaggle에서 제공된 4종류의 motor imagery EEG 측정 데이터가 train 데이터와 test 데이터로 나누어 제공되었다.

2.1 데이터의 형태

데이터는 'input'과 'label' 두 개의 키를 가진 딕셔너리 형태로 제공되었다.

2.1.1. input

Input 데이터의 형태는 (4608, 1, 22, 1125)이다. 이는 1125의 시간 동안 22개의 전극에서 측정

된 4608개의 샘플 데이터를 의미한다.

2.1.2. label

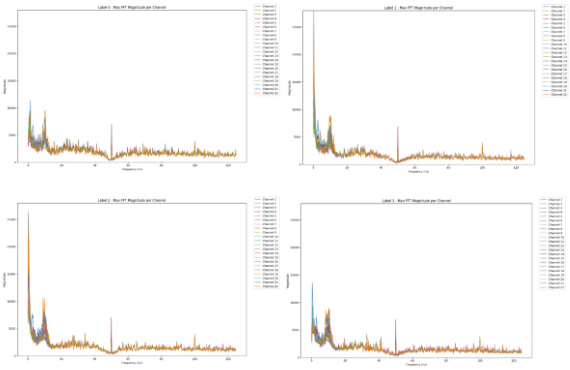
Label 데이터는 [1, 0, 0, 0], [0, 1, 0, 0], [0, 0, 1, 0], [0, 0, 0, 1] 4종류의 라벨 데이터가 제공되었다. 원-핫 인코딩 형태로, 각각 1152개씩 있는 균형 데이터셋이다.

3. 분석 방법

3.1. 밴드파워의 차이를 이용한 분류

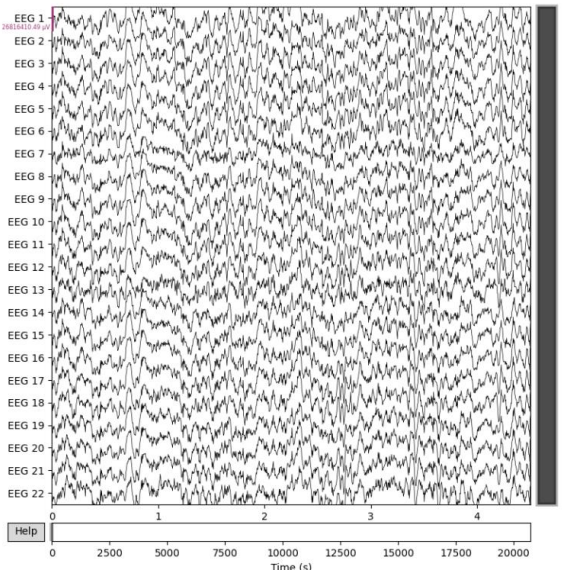
3.1.1. EDA 과정

원-핫 인코딩 형태로 주어진 라벨 데이터를 0, 1, 2, 3의 형태로 변환했다. 이후 라벨에 따른 데이터의 차이를 알아보기 위해서 각각의 라벨을 푸리에 변환 시킨 결과를 시각화해서 살펴보았다.



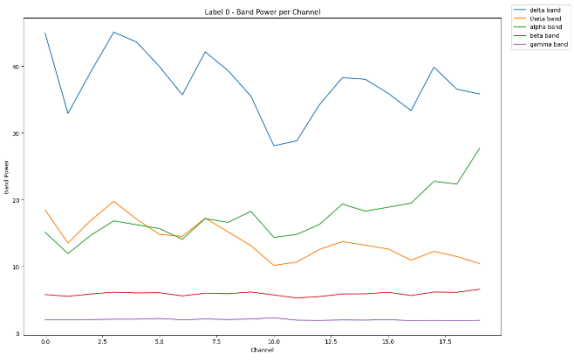
3.1.2 Bad Channel 식별 및 제거

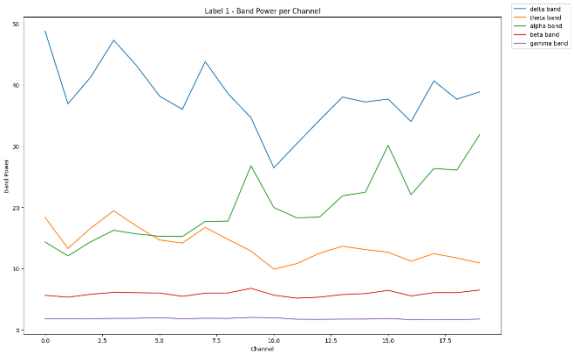
데이터의 형태를 (22, 4608, 1125)로 변환한 후 Raw 데이터를 모든 채널에 대해 PSD plot으로 나타냈을 때, 모든 채널이 육안 상으로는 거의 차이가 없었지만 7, 8번 채널이 특정 구간에서 거의 변동이 없는 구간이 있어 bad channel로 인식하고 삭제했다.

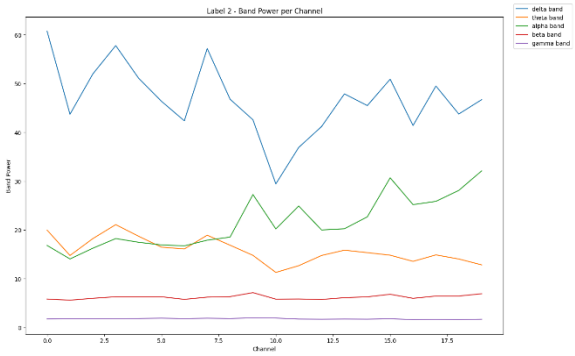


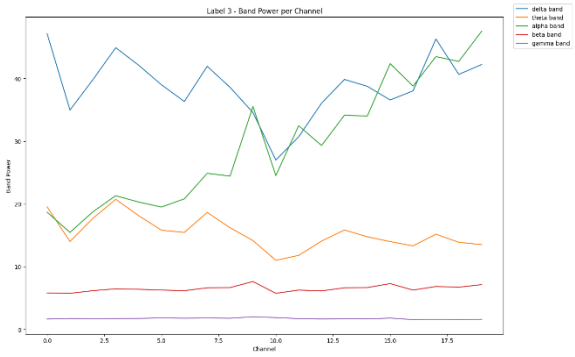
3.1.3. band power 함수 적용 및 시각화

뇌파의 주요 대역인 델타(0.5~4Hz), 세타(4~8Hz), 알파(8, 13Hz), 베타(13~30Hz), 감마(30~45Hz)를 정의하고 각 대역의 band power를 계산한 뒤 시각화했다. 채널의 각 샘플에 대해 밴드파워의 최댓값을 뽑고, 그것들의 평균을 구하는 방식으로 계산했다.









시각화 결과 세타, 베타, 감마 대역은 모든 라벨에서 거의 차이가 없었고, 델타와 알파 밴드에서 라벨별로 약간의 차이가 보여 모델링에는 이 두 개의 대역을 사용하기로 결정했다.

3.1.4. 데이터 분할 및 스케일링

Train 데이터, valid 데이터, test 데이터를 6 : 2 : 2로 분할한 후 각각의 채널에 대해 standardscaler를 활용하여 스케일링했다.

3.1.5. 대역 선별 후 band power 계산

앞서 계획한 대로 델타 대역, 알파 대역의 밴드 파워를 train 데이터, valid 데이터, test 데이터에 각각 적용했다.

도메인 자료 조사 결과 Motor imagery 분류에 주로 쓰이는 뇌파 대역은 뮤 리듬과 베타파라는 것을 알게 되었다. 뮤 리듬의 경우 8~14Hz로 알파파와 같아 그대로 적용했고, 베타파의 경우 시각화 결과에서는 라벨에 따른 차이를 확인할 수 없어 사용하지 않았다.

Band power 계산 후 각각의 데이터셋의 shape는 다음과 같다.

X_train_bandpower	(3686, 20, 2)
X_valid_bandpower	(922, 20, 2)
X_test_bandpower	(576, 20, 2)

3.2. Modeling

모델링 과정에서는 CNN을 사용하여 각각의 라벨을 분류했다.

3.2.1. CNN 모델 정의

입력 데이터 형식이 (3686, 20, 2)로 3차원 텐서인데, conv2d 레이어는 4차원 텐서를 입력으로 받기 때문에 새로운 차원을 추가하여 4차원 텐서로 변환했다.

컨볼루션 레이어 3개, pooling layer 3개, fully connected layer 3개, dropout layer 1개를 사용했다. Activation function으로는 ReLU를 사용했다.

Input
3*2 conv, 32, padding=1, pooling(kernel_size=(2, 1), stride=(2, 1), no padding
3*1 conv, 64, padding=1, pooling(kernel_size=(2, 1), stride=(2, 1), no padding
3*1 conv, 128, padding=1, pooling(kernel_size=(2, 1), stride=(2, 1), no padding
fc, 128
Dropout(0.5)
fc, 64
fc, 4

output

3.2.2. 손실 함수 및 옵티마이저

손실 함수로 교차엔트로피를 사용했고 학습률 0.001의 Adam을 옵티마이저로 설정했다.

3.2.3. 모델 훈련 및 평가

Epoch을 20으로 설정하고 train loss, valid loss, valid accuracy을 비교했다. 20 epoch 후 train loss: 1.2788, valid loss: 1.3204, valid accuracy: 0.3547라는 결과를 얻었다.

3.3. test 데이터에 적용 및 결과

완성한 모델에 test 데이터를 적용한 결과

class	데이터 개수
0	146
1	33
2	261
3	136

위와 같이 클래스가 분류됐다. 캐글 제출 점수는 0.2934로 좋지 못한 성능을 보였다.

3.4. 개선할 점

Bad channel을 육안으로 식별하기 어려웠는데, 명확한 근거 없이 bad channel을 선정한 것 같다. 또한 델타 대역, 알파 대역의 밴드 파워가 라벨 0, 1, 2와 3 사이에는 차이가 존재했지만, 라벨 0, 1, 2 사이에서는 거의 유사하게 움직인 것 같아 분류가 잘 되지 않은 것 같다. 각 라벨별로 다르게 움직이는 특징을 찾아서 이용해야겠다고 생각했다.

3.5. 분석 방법 변화 및 결과 개선

3.5.1. Continuous Wavelet Transformation (CWT) 계산

CWT는 신호의 시간-주파수 분석에 사용되는 방법이다. CWT를 이용한 기술은 신호의 다양한 스케일을 웨이블릿으로 분해하여 신호의 시간-주파수 표현을 제공한다.

연속 웨이블릿의 표현수식은 다음과 같다.

$$w_s(\alpha,\tau) = \alpha^{\frac{1}{2}} \int s(t)\phi\left(\frac{t-\tau}{\alpha}\right)dt$$

$s(t)$는 입력 신호이고, α는 웨이블릿 변환의 스케일링이고 ϕ는 웨이블릿 기반 함수, τ는 시간 오프셋이다. 여러 웨이블릿 기반 함수가 존재하지만 여기에서는 Morlet 웨이블릿을 선택하였다.

3.5.2. Power Spectral density(PSD)

Power Spectral density(PSD)는 신호의 frequency domain 에서 각 주파수 별로 차지하는 power 를 나타낸 것이다. 샘플 주파수는 256, 세그먼트 길이는 500 으로 실행하였다. psd 적용 시에는 fft 가 같이 진행되므로 별도의 푸리에 변환 과정을 거치지 않았다.

psd 적용 후 (4608, 22, 1125)의 형태의 데이터가 (4608, 22, 126)의 형태로 변환되었다.

3.5.3. CNN 모델 생성

Activation function은 출력 레이어에서는 softmax, 나머지 부분들에서는 ReLU를 사용했다. 구조는 다음과 같다.

Input(22, 126, 1)
3*3 conv, 32

pooling(kernel_size=(2, 2), stride=(2, 2), no padding Dropout(0.25)
3*3 conv, 64 pooling(kernel_size=(2, 2), stride=(2, 2), no padding Dropout(0.25)
3*3 conv, 128 pooling(kernel_size=(2, 2), stride=(2, 2), no padding Dropout(0.25)
fc, 128 Dropout(0.5)
fc, 64 Dropout(0.5)
fc, 4 Dropout(0.5)
output

3.5.4. 손실 함수 및 옵티마이저

손실 함수로 교차엔트로피를 사용했고 학습률 0.001의 Adam을 옵티마이저로 설정했다.

3.5.5. 모델 훈련 및 평가

psd외의 다른 전처리를 따로 거치지 않고 학습시켰다. 또한 batch 사이즈를 64부터 서서히 줄여 나갔는데 batch 사이즈가 8일 때 성능이 가장 좋았다.

Epoch을 50으로 설정하고 train loss, valid loss, valid accuracy을 비교했다. 50 epoch 후 valid loss: 1.2983, valid accuracy: 0.4154라는 결과를 얻었다.

의 결과가 나왔고 캐글 결과는 39.583%가 나와 성능이 꽤 향상 되었지만 여전히 높은 성능을 내지는 못했다.

3.6. test 데이터에 개선 모델 적용

캐글 제출 점수는 39.583%로 앞서 받은 점수인 29.34%보다 약 10% 증가했다.

4. 참고문헌

[1]https://youtu.be/Mc9PHZ3H36M?si=ktE-nDG1s1krq2x2

[2] David Lee, Hee-Jae Lee, Sang-Goog Lee. (2014). Motor Imagery EEG Classification Method using EMD and FFT

[3] Sang-Hoon Park, Sang-Goog Lee.(2015). A Method of Feature Extraction on Motor Imagery EEG Using FLD and PCA Based on Sub-Band CSP