

방학세미나 후기

정해줘요 학회장팀

권가민 방건우

INDEX

1. 출제 의도
2. 학회장팀이 짚고 싶은 점
3. 피드백
4. 1등팀 발표

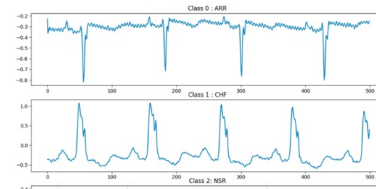
1

출제 의도

DATA

ECG_MITBIH-BIDMC

A 3-Class ECG Dataset



PhysioNET Repository에 구축된 심장 박동 데이터

여러 신호로부터 환자의 심장병 여부 예측

0 : 정상 / 1 : 부정맥 / 2 : 심부전

목적

불균형 클래스에 대한 접근 연습

강한 노이즈 데이터에 대한 노이즈 처리 접근

도메인에 맞는 전처리와 모델 선택 방법 연습

평가지표

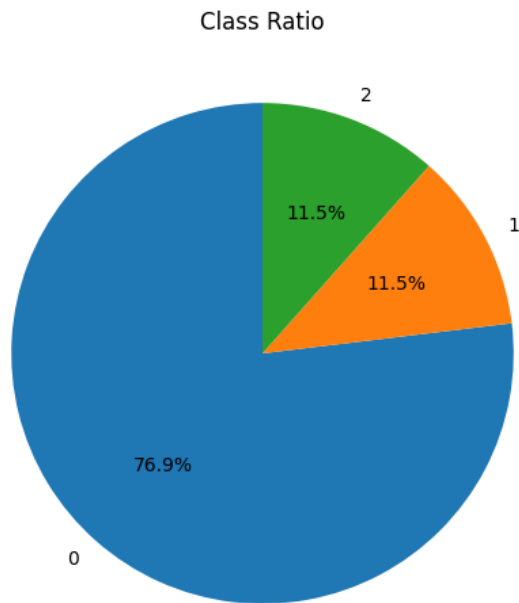
Score Function

: 클래스마다 다른 점수(score) 발생

$$\begin{aligned} \text{Score Function} = \\ 1 \times \text{TP}(\text{Class 0}) + 10 \times \text{TP}(\text{Class 1}) + 11 \times \text{TP}(\text{Class 2}) \end{aligned}$$

분류 모델의 성능을 평가하는 지표로 사용

평가지표





2팀 감사해요 ㅎㅎ

클래스의 분포가 약 20:3:3인
Imbalanced Data의 분류 문제

불균형 데이터에 대한 성능 평가에는
 $n = 30$ 일반적으로 Accuracy보다는
F1 Score 등의 평가지표를 사용함

분류 모델의 성능을 평가하는 지표로 사용

평가지표

	실제로 암에 걸린 경우	실제로 암에 걸리지 않은 경우
암으로 진단	제대로 진단하였음	<p>(예시) 병원비를 낭비함, 시간을 낭비함 등등 ...</p>  <p>그래도 건강해서 다행이다!</p>
암으로 진단하지 않음	<p>(예시) 치료를 하지 못해 암이 크게 번짐</p> 	제대로 진단하였음

위 예시처럼, 회사가 파산하는 경우는 **흔치 않지만 예측 실패 시 비용 ↑**

따라서 부정맥, 심부전을 탐지하였을 때의 점수를 높게 설정한
Score Function을 통해 성능 평가

분석 과제

불균형 클래스

0과 1,2가 20:3:3의 비율로 분포
Minor Class에 대한 예측

모델 선택 및 과적합 방지

데이터의 특성에 따른
적절한 모델 선택 및 과적합 방지

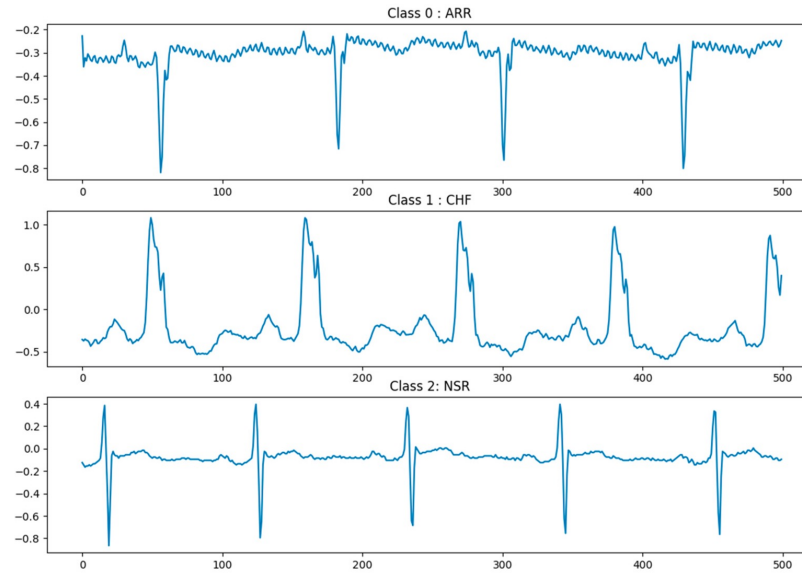
NA값 처리

NA값을 대체 또는 제외 필요
EDA를 통해 적절한 대체 필요

노이즈 처리 문제

범주에 따른 강한 노이즈
적절한 대처가 필요

결측 처리



원본 데이터는 결측치와 노이즈가 없는 문제였음

→ 죄송해요!!! 우하하...

여러분들이 얼마나 신호 데이터를 잘 처리하는지 한번 확인해봤습니다

(대신 Year, S/N, Country로 힌트 줬어요 ㅎㅎ)

2

학회장팀이 짚고 싶은 점

결측치 보간

다들 잘해서 패스함! ㅎㅎ

**FFT**

Fast Fourier Transform 기반
사인파를 통해 나타나기 때문에
갑작스러운 변화 신호에 대응이 불가함

Wavelet Transform interpolation

여러 Wavelet 함수를 이용한 보간
Small scale factor를 사용하여
갑작스러운 신호에 대응이 가능함

변수 변환

사용하는 모델의 계열에 따라서 변수 변환의 차이가 발생할 수 있음



머신러닝 방법

Tsfresh나 주기 기반 분석을 통해
열 단위 변수들로 변환하여 사용
(트리 모델은 열 단위에 최적화)

딥러닝 방법

변수 변환에 큰 관계 X
그냥 스케일링에 잘 신경쓰자!

스케일링

신호데이터에서 마음대로 스케일링은 위험!

하지만, 범주 간 차이가 존재한다면 스케일링을 고려하거나, 모델을 n 개 설계해야 함



행 단위 표준화보다는, 범주 단위 표준화가 더욱 적합할 수 있다!

문제에서 S/N 앞부분 태그 별로 분포가 다름을 확인해야 함

불균형 처리

트리 모델의 cost sensitive 처리나, imbalance loss 설계만으로는 부족함
(극단적인 분포에 대해서는 원론적인 대처가 필요)

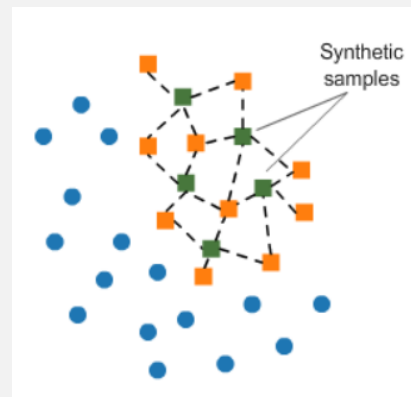


Oversampling은 거의 필수! 얼마나 Sampling 할 것인지는 잘 선택하기!

그러나, SMOTE류는 열 단위 거리 기반이므로

시계열 특성을 고려하지 못함.

ROS나 TS용 Sampling 사용이 권장됨
(GAN도 good!)



불균형 처리

트리 모델의 cost sensitive 처리나, imbalance loss 설계만으로는 부족함
(극단적인 분포에 대해서는 원론적인 대처가 필요)



Tsfresh로 시계열 변수를 생성했다면 조금 다른 양상이 가능함

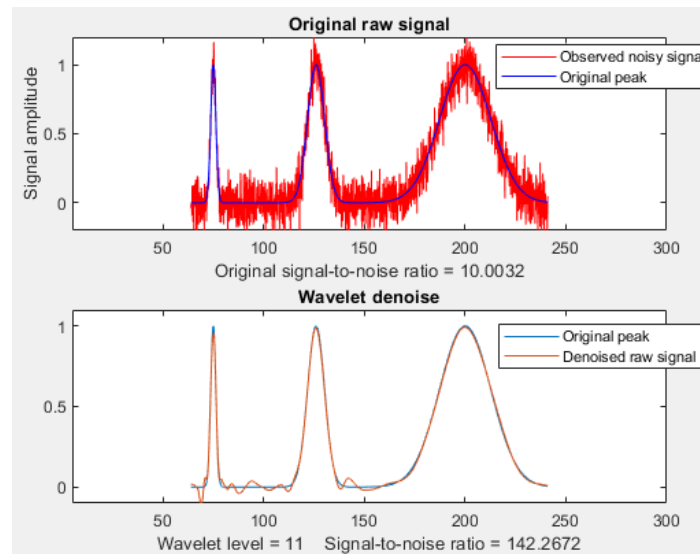
열 기반 특성으로 변환되었으므로

SMOTE, BorderlineSMOTE, SMOTE-ENN, SMOTE-Tomek, Adasyn

시도해볼 수 있음

노이즈 처리

노이즈가 많이 분포되어 있을 수록 peak를 잘못 찾아 주기가 망가질 수 있음



FFT, wavelet 기반 노이즈 처리, 시계열 분해 후 주기만 사용 등 고려할 수 있음

Score Function

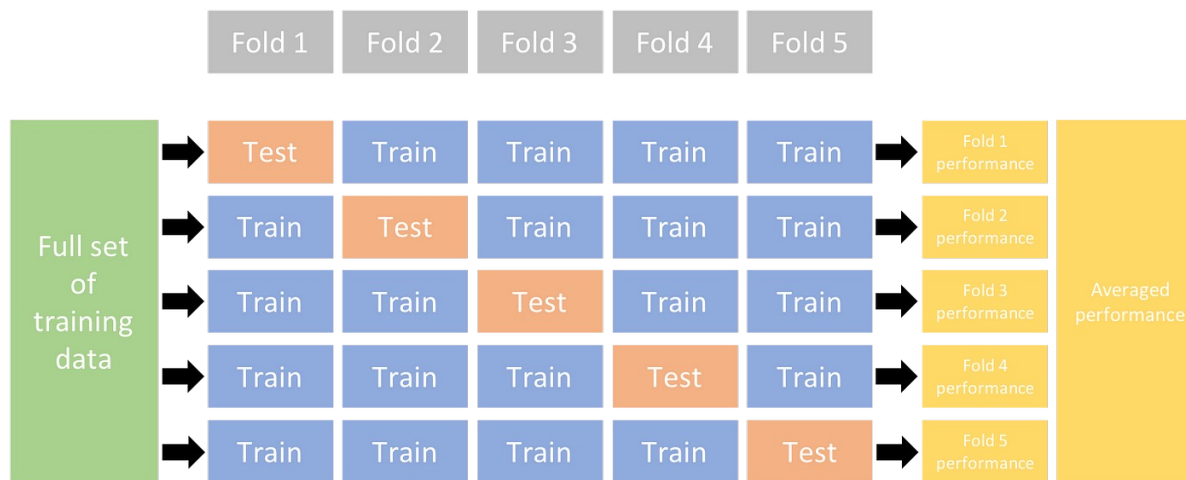
Accuracy, F1-Score등으로 성능을 어림짐작하는 것은 위험한 선택



반드시 문제에서 주어진 Score-Function을 통해서 optuna 구성할 것

Validation

랜덤 시드가 고정된 상황에서 단순 holdout validation을 신뢰할 것인가?



과적합 방지를 위해 K-CV, LooCV등 다양한 전략을 사용하여 모델 구성하기

모델 선택

항상 머신러닝 or 딥러닝 중 택1 해야하는 것은 아님!

Model Name	YearPrediction	MSLR	Epsilon	Shrutime	Blastchar
XGBoost	77.98 ± 0.11	$55.43 \pm 2e-2$	$11.12 \pm 3e-2$	13.82 ± 0.19	20.39 ± 0.21
NODE	76.39 ± 0.13	$55.72 \pm 3e-2$	$10.39 \pm 1e-2$	14.61 ± 0.10	21.40 ± 0.25
DNF-Net	81.21 ± 0.18	$56.83 \pm 3e-2$	$12.23 \pm 4e-2$	16.8 ± 0.09	27.91 ± 0.17
TabNet	83.19 ± 0.19	$56.04 \pm 1e-2$	$11.92 \pm 3e-2$	14.94 ± 0.13	23.72 ± 0.19
1D-CNN	78.94 ± 0.14	$55.97 \pm 4e-2$	$11.08 \pm 6e-2$	15.31 ± 0.16	24.68 ± 0.22
Simple Ensemble	78.01 ± 0.17	$55.46 \pm 4e-2$	$11.07 \pm 4e-2$	13.61 ± 0.14	21.18 ± 0.17
Deep Ensemble w/o XGBoost	78.99 ± 0.11	$55.59 \pm 3e-2$	$10.95 \pm 1e-2$	14.69 ± 0.11	24.25 ± 0.22
Deep Ensemble w XGBoost	76.19 ± 0.21	$55.38 \pm 1e-2$	$11.18 \pm 1e-2$	13.10 ± 0.15	20.18 ± 0.16

NODE

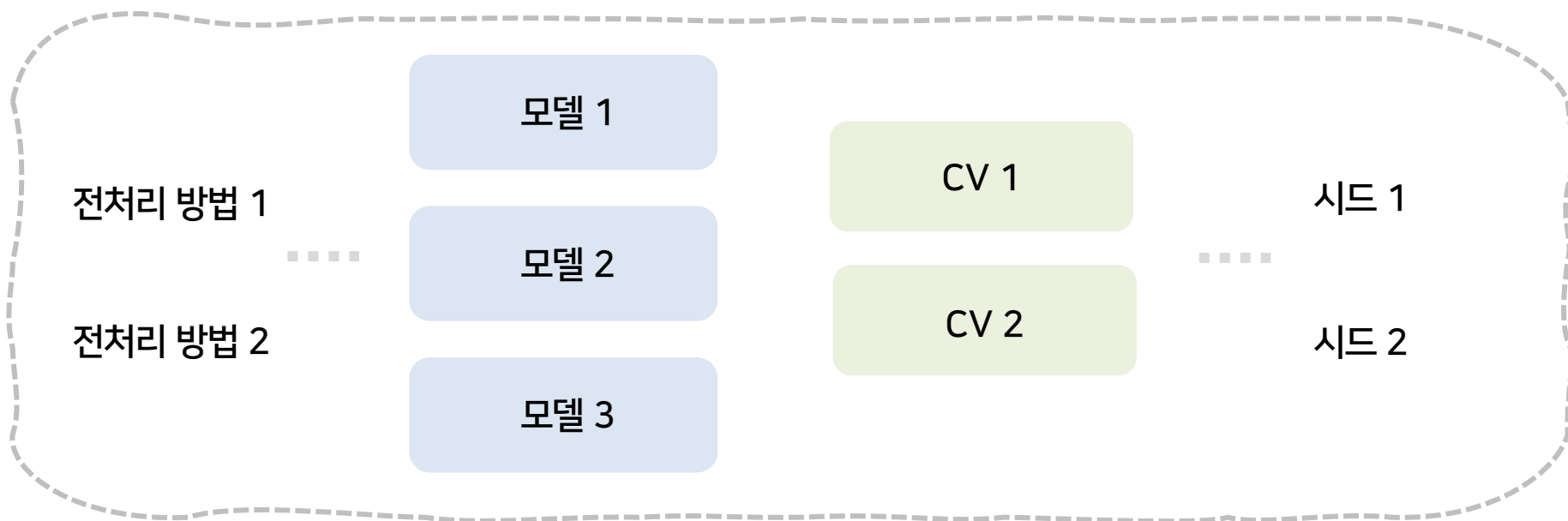
New datasets

Table 2: **Test results on tabular datasets.** Presenting the performance for each model. MSE is presented for the YearPrediction and Rossman datasets, and cross-entropy loss (with 100X factor) is presented for the other datasets. The papers that used these datasets are indicated below the table. The values are the averages of four training runs (lower value is better), along with the standard error of the mean (SEM)

DL + ML이 대표적인 좋은 앙상블의 예시임

출처 : TabNet: Attentive Interpretable Tabular Learning

앙상블



앙상블을 적용하여 **일반화 성능을 향상을 기대하여,**

Public score와 Private Score의 차이를 줄일 수 있음

불균형 데이터의 경우, 10개의 모델 중 3,4개 이상이 1인 경우 1로 두는 등

다양한 임계값 방법을 적용하여 성능을 올릴 수 있음!

3

피드백

공통

Test 데이터는 편의를 위해 Set으로 구성된 것이지만,
실제로는 개별적으로 다가오는 것입니다!

Test 데이터에 대해 결측값 대체나 scaling 등의 전처리를 진행할 경우
Test 셋에서 최댓값이나 분위수를 이용하는 것, `fit_transform()` 을 한번에 적용하는 것은
Test의 분포를 이용하는 것이므로 **Data Leakage**



만약 Test 데이터에 대한 전처리를 진행할 경우,
Train 데이터로 학습을 시킨 후 따로 모델을 대체하거나
Train 데이터에서 해당 통계량을 저장한 뒤 Test의 전처리에 사용했어야 함!

공통

마스킹된 데이터의 어려움에도 불구하고,
Label과 범주형 자료에 따른 시계열 관측치의 특성에 대해 파악하고 고민한 과정들,
NA를 처리하기 위해 여러가지 시도를 하신 과정들이 세 팀 모두 정말 인상 깊었습니다!!

또 클래스 불균형, 시계열 데이터의 특수성에 대해서도
치열하게 고민하신 과정도 잘 드러났다고 생각합니다 ㅎㅎ
앞으로 모델링할 때 방학 세미나 경험이 많은 도움이 되었으면 좋겠습니다~!~!

코드 역시 마크다운을 이용해 깔끔히 정리해주신 덕분에 채점이 편했어요 ㅎㅎ 감사합니다.
그리고 무엇보다 같은 기수끼리 친해지자는 목적에서 보너스 점수를 넣었는데,
팀 구분 없이 모두 가까워지신 거 같아 정말정말 뿌듯합니다 ^__^

4

1등 팀 발표





1팀



박상훈 권능주 곽동길 김태현 박윤아 이경미

축하드립니다~~!



다들 한 주 동안 너무
고생 많으셨습니다!

