[서식 3] 단기 산학협력 프로젝트 최종보고서

SW중심대학

산학협력 프로젝트 결과보고서

본 보고서를 SW중심대학 단기 산학협력 프로젝트 결과보고 서로 제출합니다.

(연구과제명 : MIT-BIH ArrhyThmia Dataset**을 활용한 심장질환 분류모델**)

2020. 11. 30.

참여기업 : 록스 김지현 (인)

참여학생 : 통계학과 최재혁(남) (인)

통계학과 김도희(여)

정보컴퓨터공학과 강민진(남)

참여교수 : 소프트웨어교육센터 육동철 (인)

참여인원 : 학부생 명, 대학원생 : 명, 기업체 : 명, 교수 : 명

부산대학교 소프트웨어교육센터장 귀하

산학협력 프로젝트 결과보고서[10페이지 이내]

과제명	MIT-BIH ArrhyThmia Dataset을 활용한 심장질환 분류모델			
수행기간	2020.08.24.~2020.12.04.		책임교수명	육동철 (인)
협력기관명	록스		기관담당자명	김지현
참여학생	이름	학번	팀명	록스 1팀
	팀장: 최재혁	201514150		
	팀원: 김도희	201711505		
	팀원: 강민진	201524404		
참여인원	참여교수 1명(육동철), 학부생 3명(최재혁, 김도희, 강민진), 협력기관 1명(김지현)			

추진 배경 및 목표

1. MIT-BIH 부정맥 데이터 셋 분석 : 구조 익히기

2. Beat Type Super Class Classifier 개발
- ML/DL 모델 자유롭게 적용
- 단, 모델 적용의 근거가 분명해야 함

3. 4가지 Rhythm Type Classifier 개발
- 4가지 Rhythm Type 증상에 대한 특성 분석
- 특성에 맞는 Feature 추출
- 분류에 적합한 모델 적용 및 개발

4. 성능평가
- Confusion Matrix를 활용하여 각 증상 별 분류 결과 분석

수행 결과

1. Beat Type Super Class Classifier 개발

- data: MIT-BIH Arrhythmia Database
- Target beat: N, SVEB, VEB, F, Q
- Result: F1-score/Accuracy: 98%

- 최종모델: 1D-CNN을 활용하여 개발 진행

1) window size: 252

-> 최초에 비트별 평균 길이의 분포를 구해 MEAN+2Sigma 값으로 했으나 성능이 안 좋았음.

2) Preprocessing

a. MAIN LEAD: MLII (나머지 제외)
b. 데이터 불균형으로 인해, 각각 3000으로 Over/Undersampling 진행
c. DWT + Normalization 적용

3) 1D CNN (input : 30000 * 252)

- MODEL STRUCTURE

1) conv. layer : 64 filters(6) [activation=relu]
-> batch normalization
2) MaxPooling : size=(3),strides=(2),padding="same"
3) conv. layer : 128 filters(3) [activation=relu]
-> batch normalization
4) conv. layer : 128 filters(3) [activation=relu]
-> batch normalization
5) MaxPooling : size=(2),strides=(2),padding="same"
6) conv. layer : 256 filters(3) [activation=relu]
-> batch normalization
7) conv. layer : 256 filters(3) [activation=relu]
-> batch normalization
8) MaxPooling : size=(2),strides=(2),padding="same"
9) 2 FC

- 개발 과정

1) 모델1 (03_Classification_of_ECG_signals.ipynb) (resampling=5000)

- 타겟 : N, A, V, /, L, R 대상
- 목적 : N, A, V, /, L, R 특정 비트에 대한 모델 성능 확인
- window size : 비트 길이의 mean + 2sigma = 280 + 2 * 80 = 440
(비트 길이 값(x) 정규분포를 따름. -> mean + 2sigma < x 에 해당하는 값이 97%이상을 포함)

- preprocessing
  I) 메인 리드가 MLII가 아닌 102, 104도 포함 (input data 정규화 실시하므로 포함 가능)
  II) MISSB가 많이 있던 즉, 측정이 제대로 이루어 지지 않았던 231번 제외

- 결과 : Accuracy = 95.02, F1 score = 95.03

2) 모델2 (04_Classification_of_ECG_signals.ipynb) (resampling=5000)

- 모델1과의 차이점
  MISSB가 많이 있던 즉, 측정이 제대로 이루어 지지 않았던 231번 제외 X 즉, 포함

- 결과 : Accuracy = 96.83, F1 score = 96.76

3) 모델3 (05_Classification_of_ECG_signals.ipynb): Super Class 대상2 (resampling=5000)

- 타겟 : Super class - N, SVEB, VEB, F, Q 대상
- 목적 : Super class에 대한 모델 성능 확인

- preprocessing
  I) 특정 비트가 아닌 전체 비트 모두 사용
  II) AAMI recommendation for MIT 기준 superclass 적용

- 결과 : Accuracy = 91.82, F1 score = 91.67

4) 모델4 (06_Classification_of_ECG_signals.ipynb): Super Class 대상2 (resampling=3000)

- 모델3과의 차이점
  sample 수 감소
- 결과 : Accuracy = 92.19, F1 score = 92.15

5) 모델5 (07_Classification_of_ECG_signals.ipynb): Super Class 대상3 (resampling=5000, dwt)

- 모델4와의 차이점
  Discrete wavelet trans. 적용

- 결과 : Accuracy = 95.76, F1 score = 95.64

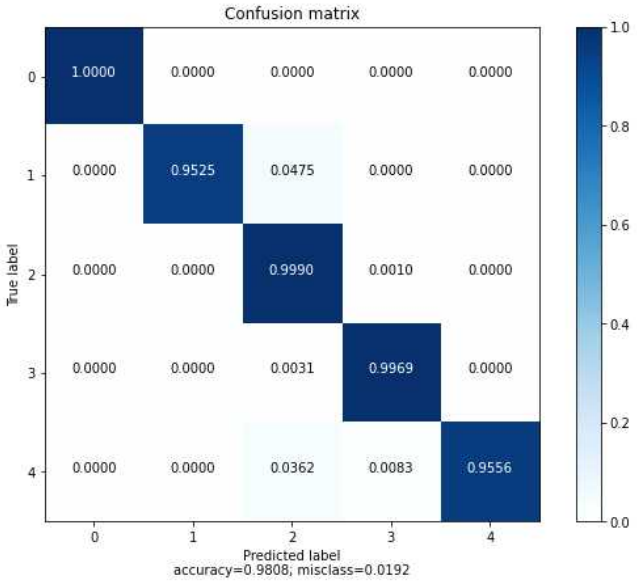
6) 모델6 (08_Classification_of_ECG_signals.ipynb): Super Class 대상4 (window size=252, resampling=5000, dwt)

- 모델5와의 차이점

window size = 252

(An Automated ECG Beat Classification System Using Deep Neural Networks with an Unsupervised Feature Extraction Technique 논문 참조)

- 결과 : Accuracy = 98.08, F1 score = 98.09



["N" , "SVEB" , "VEB" , "F" , "Q"]

2. 4가지 Rhythm Type Classifier 개발

- data: MIT-BIH Arrhythmia Database
- Target beat: "(B", "(N", "(SVTA", "(VT"
- Result: Accuracy: 81%, F1-score: 81%

- 최종결과

1) 모델: RandomForest(estimator#: 100)
Accuracy: 81%, F1-score: 81%

2) 특징
- N: PREX, AFIB와 유사한 특징을 가짐
- V(beat)의 비율이 각 Rhythm을 나누는데 가장 중요한 요소임.
-> VT, B의 경우, V(beat) 비율이 높음.
- 각 Rhythm에 포함된 beat당 R-peak 수, signal당 R-peak 수 등과 같은 R의 빈도가 중요한 요소임.
- 변수중요도: V(beat) 비율 > N(beat) 비율 > R-peak 관련 요소

- 개발과정

1) Robust detection of atrial fibrillation from short-term 기반으로 진행

: 9s 기준 Rhythm 추출해서 약 1만개 데이터를 수집했으나 실제 적용할 경우 데이터 약 100개로 사용 불가능(최초에 Rhythm 데이터가 1000개 정도)

2) 1D-CNN, 2D-CNN, ML 기반 머신러닝으로써 3가지 방식으로 접근.

:1 D-CNN/2D-CNN: 특징 추출 불가 및 데이터가 매우 적었기에 실패.

I) 타겟 클래스에 대한 데이터의 경우, 약 800개로 신경망 학습에 부적합

II) 신경망의 경우, 특징을 추출하기 어려움

III) 800개의 경우는 특징만 잘 추출한다면, 고전적인 ML기법이 직관적이고 성능이 좋을 것으로 판단

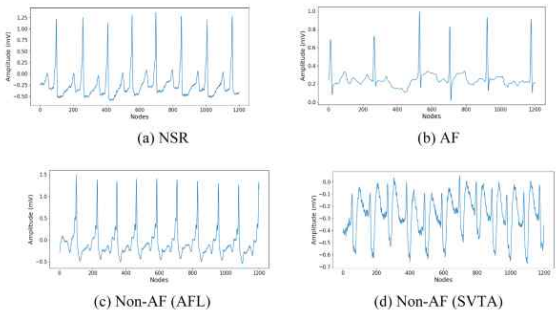
IV) Rhythm 길이가 다양하여(6만~100이하), INPUT 사이즈 정하기가 어려움.

(beat type classifier 개발 과정에서 window size 작을수록 성능이 좋은 것을 확인했으나 그러기에는 데이터 분포가 크고, 데이터는 적었음.)

3) 특징 추출

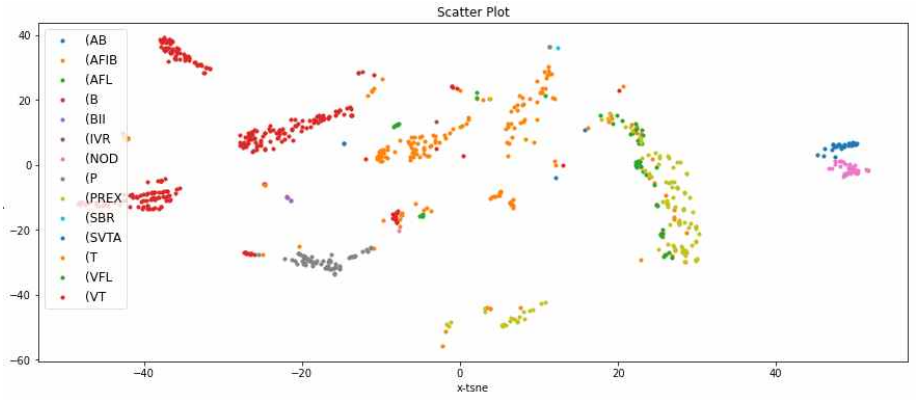
이로 인해 신호 데이터 그대로 쓸 수 없다고 판단하여 특징 추출을 실시하여 새로운 데이터를 생성을 계획함.

아래의 사진을 논문을 읽는 중 발견하고, R-Peak 간격, 개수가 Rhythm에 큰 특징이 될 수 있을 것이라 판단하여, R-peak 관련 특성과 리듬 내의 beat 종류별로 비율을 추출하여 데이터 생성함.



4) 데이터 탐색

I) t-SNE/PCA로 특징 대략적 확인



-> N의 제외한 경우, 추출된 특징으로 어느 정도 분류가 됨을 확인

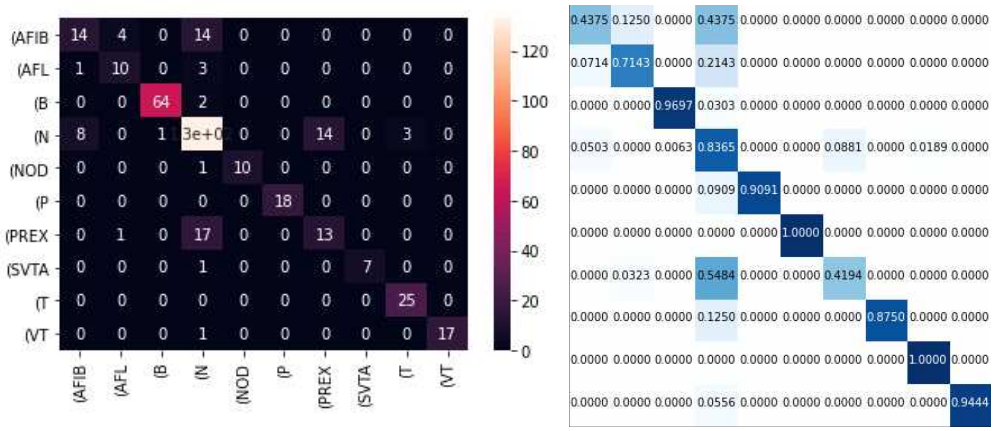
II) Decision Tree로 분류 방식 확인

: Random forest의 기본 모델이면서, 분류 로직을 확인할 수 있기에 분류방식을 대략적으로 미리 파악해보기 위함.

III) Random Forest로 분류 후 Feature Importance 확인

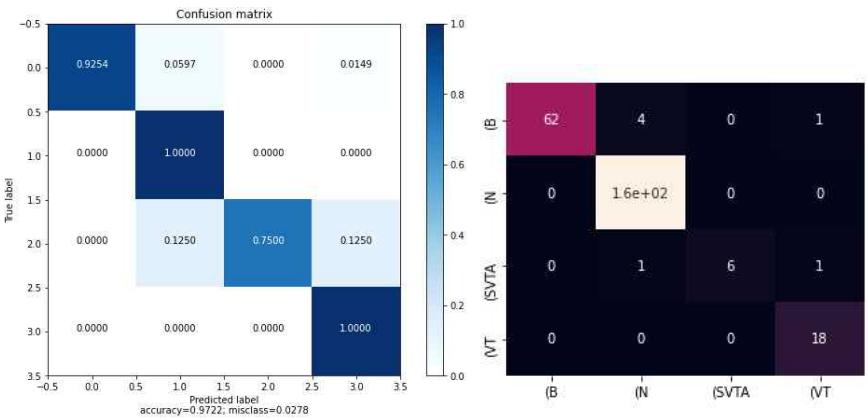
: Rhythm 내 포함된 "V", "N" 비율을 제외하고는 R peak 기반의 Feature가 주요한 분류에 영향을 미침

<전체 Rhythm 타입 적용>



- 결과 : Accuracy = 81%, F1 score = 81%

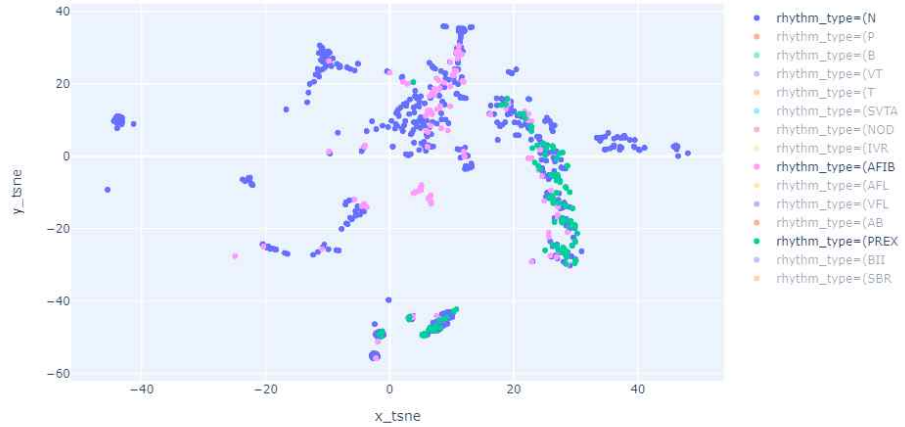
<타겟 Rhythm 타입 적용>



- 결과 : Accuracy = 97%, F1 score = 97%

두 경우 모두, 예측율 70%대 이하로 떨어지는 것은 데이터 부족으로 인한 것으로 판단

- 특히 예측율 40% 이하인 PREX, AFIB의 경우, N의 분포 내에 대부부분 겹치는 것을 확인



과제 성과 및 향후 계획

1. 과제 성과

1) Beat Type Super Class Classifier 개발: Accuracy = 98.08, F1 score = 98.09
2) 4가지 Rhythm Type Classifier 개발: Accuracy = 97%, F1 score = 97%

2. 향후 계획

- 이번 기회를 통해서 신경망과 이에 관련된 기법들에 깊게 배우고 적용해볼 수 있었습니다.
- 마지막 과업 때, 실패했던 2D-CNN을 이후에 다시 한 번 더 사용해볼 예정입니다.

< 결과 보고서 작성 주의사항 >

1. 산학협력 프로젝트 추진배경 및 목표

o 산학협력 프로젝트를 수행하게 된 배경 및 필요성, 중요성 등을 구체적으로 기술

2. 산학협력 프로젝트의 수행결과

o 프로젝트 수행 방법과 결과 내용을 구체적으로 기술

o 프로젝트 결과물 제출

- 사용자 UI & 관리자 UI이 있을 경우 화면 캡쳐 제출
- 실행되는 사이트 URL, 앱 apk 다운로드 URL제출
- 데이터베이스 사용을 했을 경우 테이블 명세서 제출
- 인공지능 학습에 해당되는 경우 학습 모델 관련 내용과 학습 데이터 일부 제출

o 프로젝트 수행 내용과 관련한 사진, 문헌(논문, 보고서 등), 데이터, 인증서, 제품 등을 포함

3. 과제성과 및 향후 계획

o 산학협력 프로젝트 성과 및 향후 계획 내용을 기술