연구논문/작품 제안서

2021 년도 제 1 학기

논문/작품	○논문( ) ○작품(o) ※ 해당란에 체크
제목	표본 코호트 및 BMC Medical Informatics and Decision Making 기반 심부전증 예측 모델
GitHub URL	https://github.com/ggume/forgrad
팀원명단	최 서 연 (인) (학번: 2015312653)

2021 년 03 월 18 일

지도교수 : 고 영 중 서명

1. 과제의 필요성

1.1. Abstract

심혈관질환(Cardiovascular diseases, CVDs)은 전 세계적인 사망원인 1위로, 매년 약 1790만 명이 사망하는데, 이는 전체 사망자의 31%를 차지한다. 심혈관 질환은 뇌와 심장에 생기는 병으로, 혈관에 이상이 생기는 경우까지 포괄한다. 이 질환은 암에 이어 한국인의 질병 사망원인 2순위를 차지하는 질병이지만, 조기에 발견해낼 수만 있다면 치료의 예후는 굉장히 빠른 편이다. 따라서, '조기 예측 및 진단'이 아주 중요한 질병 중에 하나라고 할 수 있다. 실제로 지난 수십 년 동안 심혈관계 질환에 의한 사망률은 세계적으로 감소하고 있는 추세인데, 국내에서는 오히려 사망률이 늘어나고 있는 것을 확인할 수 있었다.

본 과제에서는 심부전(Heart Failure)에 초점을 맞추어 국민건강보험공단 건강검진 코호트 데이터베이스와 2020년에 배포된 BMC Medical Informatics and Decision Making의 자료를 이용해 그를 예측하는 모델을 여러 기법을 이용하여 예측해보고, 비교 및 평가 할 것이다.

1.2. 서론

심혈관질환(Cardiovascular diseases, CVDs)은 전 세계적인 사망원인 1위로, 매년 약 1790만 명이 사망하는데, 이는 전체 사망자의 31%를 차지한다. 그중 심부전(heart failure)은 CVD에 의해 야기되는 가장 일반적인 질환인데, 심장의 구조적/기능적 이상으로 인해 심장이 혈액을 받아들이는 이완, 수축 기능이 감소하여 신체 조직에 필요한 혈액을 제대로 공급하지 못해 발생하는 질환군을 통칭한다. 2017년 기준, 8명 중 1명의 사망원인이며, 이 질환으로 인한 사회적 비용은 307억 달러에 육박한다.(2012년 기준) 매우 다양한 원인에 의해 심부전이 초래되는데, 심장 혈관 질환 등이 가장 흔한 원인이며 장기간의 빈맥(빠른 맥박), 지속적인 과도한 음주, 극심한 스트레스 등 또한 원인이 될 수 있다.

대부분의 심혈관계 질환은 흡연, 식습관, 비만, 운동 부족, 음주 등을 지속적으로 관리한다면 예방이 가능하며 일단 진단을 받은 뒤에도 생활 습관 개선을 통해 병증이 완화될 수 있다. 따라서, 심혈관 질환을 앓고 있거나 심혈관 위험도가 높은 사람 중, 심부전이 발생할 확률이 높은 사람들에 대하여 조기 발견 및 관리를 통해 사망률을 낮출 수 있는 한편 사회적 비용 또한 감소하길 기대할 수 있을 것이다. 우리는 심부전의 각종 지표들을 가진

사람들의 질병으로 인한 사망여부를 분류(classification)해 보고 여기에 사용된 특징 요소들을 통해 심부전으로 인한 사망을 예측한다.

2. 선행연구 및 기술현황

이미 의료 데이터를 활용한 연구들은 다양하게 진행되고 있다. 선행 연구는 두 파트로 나누어서 살펴보도록 한다.

2.1. 국내외 심혈관질환 위험도 평가 방법

2.1.1. Framingham risk score

Framingham heart study는 동맥 경화성, 고혈압성 심혈관 질환의 역학 자료의 토대를 마련한 대표적인 연구로 모든 형태의 심혈관 질환의 유병률을 파악하기 위해 1948년에 시작되었다. 당시 5000여명의 대상자들이 2년 간격으로 건강 진단을 20년간 지속하였으며 현재 3대에 걸쳐 연구 중이다. 오랜 기간에 걸쳐 전향적인 역학조사를 하며 심혈관 질환의 발병률에 대하여 영향을 미치는 인자들과 개개인의 위험 요소를 어느정도는 정량적으로 계산할 수 있게 되었다. 1976년부터 heart dease prediction model이 구축되었는데, 이후 점차 연령을 확대하고 성별, 콜레스트롤 지수, 혈압, 당뇨, 고혈압 등을 중요한 인자로 지정하여 지금의 모형을 만들어내었다. 이는 수많은 심형관질환 연구의 가이드 라인이 되었다.

성별, 연령, 흡연, 수축기 혈압, 총 콜레스테롤, HDL-콜레스테롤 등을 점수화하여 향후 10년간의 심장병 발생 위험을 예측한다. 그러나 여러 후속 연구에서, 이 연구가 실제 위험도에 비해 위험도를 과다하게 측정하는 경향이 있다고 말한다. 실제로 우리나라에서는 3배, 중국에서는 5배, 그리고 독일에서도 2배가량 과대 추정하는 경향을 보이는 것을 여러 논문에서 지적하였다.

2.1.2. Korean Heart Study 모형(KHS모형)

우리나라에서 개발된 Framingham 방식을 채용한 심장질환 예측 모형이다. 1996년부터 2001년까지 18개의 검진센터에서 종합검진을 받은 30-74세의 성인 268,315명을 대상으로 추적 관찰을 진행하였다. 이 연구에서는 나이, 혈압, 총콜레스테롤, 흡연, 당뇨병여부를 기본적인 예측인자로 삼고 고밀

도콜레스테롤과 저밀도콜레스테롤, 중성지방까지도 모델에 포함시켜 정확도를 높였다.

2.2. 연구 기술

의료 데이터 연구에서는 Electrocardiogram(심전도, ECG) 데이터를 이용하는 연구가 주류를 이루고 있다. 2017년에는 발작성 심박세동 환자를 선별하는 연구에서는 ECG 데이터를 기반으로 Convolutional NeuralNetwork(합성곱신경망, CNN) 모델을 이용하였고, 같은 해의 다른 연구에서는 ECG 신호를 전처리하여 VGG16 모델을 이용하였다. 2018년에는 Detecting atrial fibrillation by deep convolutional neural networks에서 ECG 데이터를 바탕으로 Deep Convolutional Neural Network(DCNN)을 이용하였다.

3. 작품 전체 진행계획 및 구성

3.1. 데이터 정제

3.1.1. 데이터

본 연구에서 Heart Failure를 예측은 두가지 방면으로 진행된다. 첫째는 국내의 심장질환, 그중 특히 Heart Failure를 예측하기 위하여 국민건강보험공단 건강검진 코호트 데이터베이스를 이용한다. 건강검진 코호트 데이터베이스는 건강검진 수검자 중심의 의료 이용 및 건강 결과 분석을 위해 제공되는 연구용 DB이다. 본 연구에서는 단순무작위 추출을 통해 표본을 선택하고 건강검진결과 및 의료용 정보를 구축할 예정이다. 둘째는 국내가 아닌 세계적 동향 자료로 예측을 수행하기 위해 미국 BMC Medical Informatics and Decision Making에서 2020년에 David Chicco, Giuseppe Jurman이 배포한 자료를 활용한다.

3.1.2. Feature 추출

본 연구에서는 나이, 빈혈 여부, CK(크레아틴키네이스)수치, 당뇨 여부, 심장 박출률, 고혈압 여부, 혈소판 수지, 혈청 크레아티닌, 혈청 나트륨 수치, 성별, 흡연 여부, 식전혈당, 체질량지수(BMI), 수축기 혈압, 이완기 혈압, 혈색소, 연간 흡연량, 운동의 여부 등의 Feature들 중에 심장병 예측에 유용할 것이라고 생각되는 Feature들을 추출 할 예정이다. 특징요소 추출을 위해 sklearn의 ExtraTreesClassifiet를 사용하여 중요도를 파악할 것이다. 앞서

언급한 Feature들은 이미 심혈관 질환을 예측하는 데에 중요한 지표로 잘 알려져 있는 것들인데, 생소한 것들을 살펴보면 다음과 같다.

CK수치는 근육의 염증상태를 나타내는 지표이다. 운동을 심하게 했을 때, 심근경색 등 근육의 손상이 생길 때 CK농도가 상승한다. 병원에서 심근경색의 위험이 있어 보일 때, 환자의 CK수치를 확인하는데 이 정보는 골격근, 심장, 또는 뇌 중 어떤 곳이 손상되었는지 결정하는 데에 도움을 준다.

심장박출률(Ejection Fraction, EF)은 심장이 박동할 때마다 심장에서 박출되는 혈액의 비율로, 심장 기능의 중요한 척도로써 활용되고 있다. EF는 55%~70% 사이일 경우 정상범주에 속한다. 심부전이 의심되는 환자의 경우 좌심실 박출률이 40%미만이다. 치료가 얼마나 효과가 있는지에 따라 EF는 오르거나 내릴 수 있다.

크레아티닌은 근육에서 포그포크레아틴의 분해 산물이며 일반적으로 신체에 의해 상장히 일정한 속도로 생성된다. 혈청 크레아티닌은 콩팥에 의해 변하지 않고 배설되는 근육 대사의 부산물로 굉장히 쉽게 측정되는 지표 중 하나이며, 콩팥 건강의 중요한 지표로서 활용되고 있다. 혈청 크레아티닌의 기준 간격은 0.6~1.3mg/dL(53~115 μmol/L)로 이 수치에 이상이 있을 경우 이미 늦은 진단이라고 할 수 있다.

혈청 나트륨의 정상 수치는 135~145mEq/L로, 수치가 높다면, 뇌, 신경근, 심혈관계의 이상을 생각해볼 수 있다

3.2. 학습 모델 선택

연구를 진행하면서 사용할 학습 모델을 살펴본다. 실험 진행시 결과값이 계속 달라지는 것을 방지하기 위하여 시드를 고정시킬 예정이다. Python 환경에서, scikit-learn을 활용할 예정이며 추후 변경의 여지가 있음을 명기한다.

3.2.1. Logistic Regression(로지스틱 회귀)

로지스틱 회귀는 분류 문제에서 이미 아주 보편적으로 사용되고 있는 모델이다. 독립변수의 선형 결합을 이용해 종속 변수의 값을 예측하는 모델로, 수치 값으로 구석된 feature vector인 X, 확률 값으로 정의되는 종속 변수 y 사이의 관계가 선형적으로 정의됨을 가정하고 그것을 잘 표현할 수 있는 계수를 추정하는 모델이다.

3.2.2. Decision Tree(의사결정나무)

Decision Tree를 활용한 분석은 rule을 학습을 통해 자동으로 찾아내어 나무 구조로 도표화하여 classification과 prediction을 수행하는 분석 방법이다. 반복해서 학습할 수 있는 것이 장점이자 단점인데, 정확도를 높일 수도 있지만 그만큼 과적합 문제도 빈번하게 발생하기 때문이다.

3.2.3. Random Forest(랜덤 포레스트)

특징 요소를 추출하기 위하여 사용하는 방식은 여러 개가 있지만, EXTRA TREES CLASSIFIER를 사용한다. 이 classifier는 앙상블 학습의 일부인데, 앙상블 학습 기법은 sklearn 라이브러리에 이미 구현이 되어있어 사용하기 쉬울 뿐더러 비교적 빨라 널리 사용되는 기법이다. 앙상블 기법은 랜덤 포레스트 알고리즘을 사용하는데 랜덤 포레스트는 각 독립변수의 중요도를 계산할 수 있다는 장점이 있다. Random forest는 결정 트리에 기반하여 만들어진 모델으로 여러 개의 결정트리 classifier가 생성이 되고 각자의 방식으로 데이터를 샘플링하여 개별 학습 후 최종적 voting을 통해 예측을 수행하는 기법이다. 포레스트 안에서 사용된 모든 노드에 대해서 어떤 독립 변수를 사용하였는지를 알 수 있고 imformation gain을 통해서 어떤 독립변수가 중요한지를 쉽게 알 수 있다.

본 연구에서 모을 수 있는 데이터의 feature들은 숫자가 크기 때문에 그 상관관계를 정확히 하기 위해 feature를 줄일 필요가 있는데 어느 변수가 중요한지를 빠르고 쉽게 알기 위해서 이 방법을 채택하였다. 랜덤 포레스트에서 트리를 만들 때, 노드는 무작위로 feature의 서브셋을 만들어 분할에 사용하는데 트리를 더 무작위하게 만들기 위해서 최적의 임계값을 찾는 대신 후보 특성을 사용해 무작위로 분할한 다음 그중 최상의 분할을 선택한다. 이 기법은 편향이 늘어나는 대신, 분산이 줄어드는데 이는 데이터 오류의 감소로 이어진다.

3.2.4. 기타 모델

확정적으로 어떤 모델을 사용할지 결정이 되지 않았으나 몇 가지 후보군들이 존재한다. 그 중 몇 가지를 소개한다.

Deep Neural Network(심층신경망, DNN)모델은 입력층과 출력층 사이에 은닉층으로 이루어진 인공신경망으로, 독립 변수들과 종속 변수 사이의 관계를 비선형 함수로 모델링이 가능하다는 장점을 갖고 있다. LightGBM 모델은 2016년 MS에서 제안한 gradient boosing 알고리즘으로 약한 분류기들을 묶어 정확도를 예측하는 방법인데 트리의 균형을 맞추지 않는 것이 특

징이다. 이 모델은 대량의 데이터를 빠르게 학습할 수 있다는 장점을 갖고 있으며 과적합에 민감해 비교적 정확도가 높다. 그 외에도 ANN, CNN, RNN과 최근 사용이 늘어나는 추세를 가진 Boost기법을 활용하는 모델 등 데이터의 특성에 따라 적용할 수 있는 모델들을 생각하고 있다.

3.3. 결과 분석 및 시각화

빅데이터를 활용한 데이터 분석이 점점 다양해지고 그 중요성을 더해가면서, 데이터를 좋은 모델로 분석해내는 능력 만큼 그것을 보기 쉽고 이해하기 편하도록 시각화하여 스토리 텔링하는 능력 또한 점차 중요해지고 있다. 따라서, 결과를 분석한 뒤 효과적인 시각화를 수행하고 이해를 위한 자료를 만들도록 한다.

3.4. 성능 평가

성능은 accuracy score와 f1 score를 주로 사용하여 평가하기로 한다. 관련 연구들을 찾아보고 더 나은 성능 평가 척도가 있다면 적용해본다.

4. 기대효과 및 개선방향

4.1. 기대효과

몇 가지의 검사 결과만으로 미리 사망에 이르기까지 할 수 있는 심부전의 가능성을 가진 사람들을 찾아내어 관리할 수 있다면 의료적으로 아주 큰 효용이 있을 것으로 예상된다. 특히나 2년을 주기로 갱신되는 표본 코호트의 데이터를 이용한다면, 대상자가 2년 뒤에 해당 질환을 가질 것인지에 대해 예측을 할 수 있을 뿐더러 2년 후에는 그 예측이 맞는지를 확인하여 연구를 한 층 더 발전시킬 수 있을 것이다. 이렇게 정확도를 올려가며 프로그램을 만든 뒤에는 건강검진을 진행한 사람들을 대상으로 어떤 질병의 위험군인지 인지시키고, 건강에 대한 경각심을 갖도록 할 수 있을 것이다.

5. 기타

5.1. 연구 일정

연구는 기본적으로 python을 이용하며, 연구 노트는 git에 주기적으로 업로드 될 것이다. 3월에는 사전 배경지식을 수집하고 공부하며, 4-5월에는 데이터의 feature을 추출하고 각 모델을 공부한다. 6-8월에는 학습한 모델들을 바탕으로 실제 데이터를 분석한다. 7-9월에는 각 모델의 성능을 분석하고 데이터 시각화 기법을 공부하며 더 나은 모델을 위한 개선점을 고민한다. 10월에는 최종적으로 fix된 모델을 통해 데이터를 시각화하고 각 모델에 대한 성능을 다각도로 분석해본다. 11월에는 최종 작품 보고서를 작성 및 마무리한다.

6. 참고문헌

[1] 국민건강보험자료 공유서비스, https://nhiss.nhis.or.kr

[2] Grundy SM, Cleeman JI, Merz CN et al. Implications of recent clinical trials for the National Cholesterol Education Program Adult Treatment Panel III guidelines. Circulation 2004;110:227-39.

[3] 안경아, 지선하 등, Korean Journal of Epidemiology Vol. 28, No. 2, Dec, 2006, 162-170

[4] 조정래, 김윤년, 이민호. "딥러닝을 이용한 심방세동 예측 연구." 대한전자공학회 학술대회 (2017): 868-870.

[5] Jee SH, Jang Y, Oh DJ, Oh BH, Lee SH, Park SW, Seung KB, Mok Y, Jung KJ, Kimm H, Yun YD, Baek SJ, Lee DC, Choi SH, Kim MJ, Sung J, Cho B, Kim ES, Yu BY, Lee TY, Kim JS, Lee YJ, Oh JK, Kim SH, Park JK, Koh SB, Park SB, Lee SY, Yoo CI, Kim MC, Kim HK, Park JS, Kim HC, Lee GJ and Woodward M. A coronary heart disease prediction model: the Korean Heart Study. BMJ Open. 2014;4:e005025.

[6] Shashikumar, Supreeth Prajwal, et al. "A deep learning approach to monitoring and detecting atrial fibrillation using wearable technology." 2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI). IEEE, 2017.

[7] Lewis, Sharon Mantik; Dirksen, Shannon Ruff; Heitkemper, Margaret M.; Bucher, Linda; Harding, Mariann. 《Medicalsurgical nursing : assessment and management of clinical problems》 9판. St. Louis,

Missouri
[8] 대한내과학회지 2015;88:127,2