중환자실 폐렴 환자에 대한 시뮬레이션 기반 시계열 사망 마커 탐지

김수현 ⁰ 이수현 고가연 안홍렬*

수원대학교 DS&ML 센터, 수원대학교 데이터과학부

{myksh0903; happyshsh72; imgayoun0}@gmail.com, hrahn@suwon.ac.kr

Simulation-Based Time Series Death Marker Detection for Pneumonia Patients in Intensive Care Unit

Suhyun Kim⁰ Suhyeon Lee Gayoun Koh Hongryul Ahn*

DS&ML Center, Division of Data Science, The University of Suwon

요 약

MIMIC-III는 중환자실 환자에 대해서 진단, 검사, 처방 등의 의료 이벤트를 시간에 따라 추적하여 데이터화한 시계열 의료 빅데이터이다. 본 연구에서는 MIMIC-III 데이터를 활용하여 중환자실 폐렴 환자에 대한 시뮬레이션 기반 시계열 사망 마커를 탐지하는 방법을 제안한다. 제안하는 방법은 폐렴 환자의 사망을 잘 예측하는 시계열 딥러닝 모델을 선정하고 의료 이벤트 값을 변형하면서 사망 위험도 변화를 분석하여 시계열 사망 마커 점수를 산출한다. 제안하는 방법을 중환자실 폐렴 환자 데이터에 적용했을 때, 사망 및 생존에 상관관계를 가지는 의료 마커 이벤트를 찾아내었다.

1. 서 론

의료 빅데이터는 4차 산업혁명 빅데이터 응용의 핵심 분야이다. MIMIC-III 데이터[1]는 공개 의료 빅데이터 중 하나로, 2001년부터 2012년까지 Beth Israel Deaconess Medical Center 중환자실에 머물렀던 46,520명 환자에 대해서 진단, 검사, 처방 등의 의료 이벤트를 시간에 따라 추적하여 데이터화한 시계열 전자의무기록(EMR) 데이터이다.

본 연구에서는 MIMIC-III 데이터를 활용하여 중환자실 폐렴(pneumonia) 환자에 대한 시계열 사망 마커를 탐지하는 방법을 제안한다. 폐렴 환자의 시계열 사망 마커란 해당 의료 이벤트가 발생했을 때, 폐렴 환자의 사망 경향이 높아지는 의료 이벤트를 의미한다. MIMIC-III 데이터에 대해서 정보이론 기반의 특성 중요도를 계산하여 의료 마커를 탐지하는 기존 연구[2, 3]가 있었지만, 정보이론 기반의 알고리즘은 비시계열 알고리즘으로, 시간에 따라 의료 사건이 발생하는 MIMIC 데이터의 시계열 특성을 고려하지 못한 한계가 있었다. 본 연구에서는 데이터 시뮬레이션을 통해 폐렴 환자의 사망에 관련한 시계열 의료 마커를 탐지하는 방법을 제안한다. 제안하는 방법으로 찾아낸 폐렴 사망 마커들로부터 폐렴의 사망 원인에 대한 새로운 단서를 발견하게 되면 폐렴의 사망률을 낮추는데 도움이 될 것으로 기대된다.

2. 문 제 정 의

2.1. 폐렴 환자 시계열 의료 데이터

폐렴 환자 시계열 데이터는 N명의 폐렴 환자에 대해서 각 환자의 사망/생존 퇴원일을 기준일(D0)로 하여 기준일을 제외한 T일 전까지 기간(D-T, ..., D-1)동안 F개의 의료 이벤트 발생을 추적한 시계열 데이터이다. [그림 1]은 어떤 날(D0)에 사망한 한 폐렴 환자에 대한 사망일 이전 10일 사이(D-10 ~ D-1)의 시계열 데이터 예시를 보여준다.

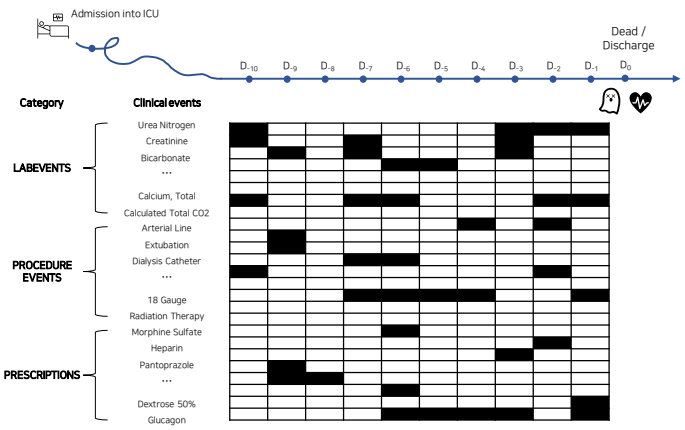

그림 1 한 환자의 10일치 의료 시계열 이벤트 데이터 예시. 가로축은 사망/퇴원일(D0)과 그 전 10일(D-10 ~ D-1)의 시점들이며, 세로축은 의료 이벤트의 종류이다. 데이터 가운데에 검은색/흰색 칸은 해당 의료 이벤트가 그 시점에서 발생/미발생 했음을 의미한다.

이 논문은 2020학년도 수원대학교 학술진흥연구비 지원에 의한 논문임

폐렴 환자 시계열 데이터는 인공지능 모델링 예측을 수행하기 위해 설명변수 데이터 X와 목적변수 데이터 Y로 구조화된다. 환자의 수를 N, 고려하는 시점(time point)의 개수를 T, 의료 이벤트 종류의 수를 F라고 하면, 설명변수 데이터 X는 해당 환자의 해당 시점에서 해당 의료 이벤트가 발생 or 미발생(1 or 0)했는지에 대한 값으로 $X = \{1,0\}^{N \times T \times F}$ 로 표현되는 3차원 행렬 데이터이다. 목적변수 데이터 Y는 해당 환자가 사망 or 생존했는지(1 or 0)에 대한 값으로 $Y = \{1,0\}^{N}$로 표현되는 1차원 벡터 데이터이다.

2.2. 폐렴 환자 사망 마커 탐지 문제

폐렴 환자 사망 마커 탐지 문제란, 폐렴 환자 시계열 데이터 X에 존재하는 F개의 의료 이벤트 $\{F_1, \dots, F_F\}$에 대해서, 폐렴 환자가 사망에 이르게 하는 영향력에 대한 중요도 점수 $\{IS(F_1), \dots, IS(F_F)\}$를 출력하는 문제이다(IS는 importance score의 약자이다). 각 이벤트의 중요도 점수로부터 상위 n개 이벤트를 선별하여 n개의 사망 마커를, 하위 n개 이벤트를 선별하여 n개의 생존 마커를 최종적으로 산출하게 된다.

3. 방 법

시계열 폐렴 위험 의료 마커 탐지 방법은 2단계(STEP 1,2) 과정으로 각 의료 이벤트에 대한 사망 마커 중요도 점수를 출력한다[그림 2].

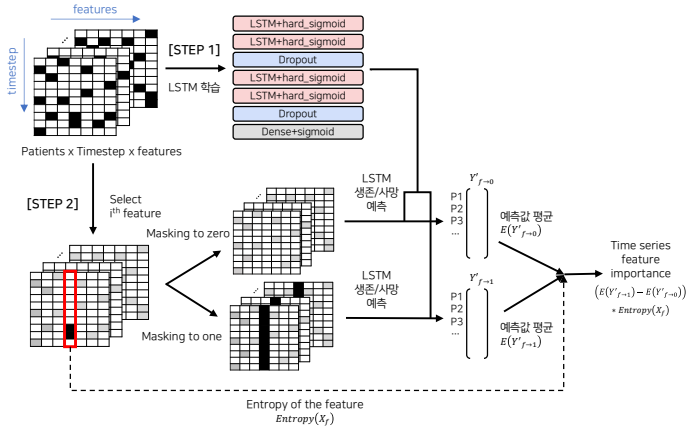

그림 2 시뮬레이션 기반 시계열 사망 마커 점수 계산 방법.

3.1 [STEP 1] 시계열 인공지능 모델 학습

첫번째 단계에서는 환자의 사망/생존을 예측하는 시계열 인공지능 모델을 학습한다[그림 2의 상단부]. 본 연구에서는 시계열 인공지능 모델로 장기 의존성 문제를 개선한 딥러닝 인공지능 모델인 LSTM(Long Short-Term Memory)을 사용하였다[4]. 손실 함수와 최적화 알고리즘으로 binary cross entropy와 adam을 사용하였으며, 최대 반복 학습 횟수(epoch)는 300회로 지정하고 과적합을 방지하기 위하여 early stopping을 적용하였다.

3.2 [STEP 2] 시뮬레이션을 통한 시계열 사망 마커 중요도 점수 계산

두번째 단계에서는 데이터를 변형하는 시뮬레이션을 통해 각 의료 이벤트에 대한 사망 마커 점수를 계산한다[그림 2의 하단부]. 사망 마커 점수를 계산하고자하는 한 의료 이벤트를 $f \in \{F_1, \dots, F_F\}$라고 할 때, 의료 이벤트 f에 대한 사망 마커 중요도 점수 $IS(f)$는 다음과 같이 계산된다.

먼저, 원본 설명변수 데이터를 X에서 의료 이벤트 f의 데이터 값을 변형함으로써, 두 종류의 시뮬레이션 데이터 $X_{f\to1}$, $X_{f\to0}$을 얻는다.

- $X_{f\to1}$: 원본 데이터 X에서 의료 이벤트 f의 모든 환자의 모든 시점에 대한 값을 발생(1) 값으로 바꾼 발생 시뮬레이션 데이터이다.
- $X_{f\to0}$: 원본 데이터 X에서 의료 이벤트 f의 모든 환자의 모든 시점에 대한 값을 미발생(0) 값으로 바꾼 미발생 시뮬레이션 데이터이다.

이렇게 발생/미발생 시뮬레이션 데이터를 생성한 뒤, STEP 1에서 원본 데이터로 학습된 LSTM 모델 M에 시뮬레이션 데이터를 예측 입력 값으로 사용하여 예측 시뮬레이션 값 $Y'_{f\to1} = M('X_{f\to1})$ 와 $Y'_{f\to0} = M('X_{f\to0})$를 얻는다.

최종적으로, 의료 이벤트 f에 대한 사망 마커 점수 $IS(f)$는 $Y'_{f\to1}$와 $Y'_{f\to0}$의 평균값의 차이와 엔트로피의 곱을 통해 계산된다.

$$IS(f) = \left(E(Y'_{f\to1}) - E(Y'_{f\to0})\right) * Entropy(X_f)$$

이렇게 정의된 사망 마커 점수는 다음과 같은 특징을 가진다.

의료 이벤트 f가 사망 마커 이벤트이면 (즉, 사망과 양의 상관관계를 가지는 이벤트이면), 사망 마커 점수 $IS(f)$는 큰 양수 값을 가진다. 이때, 발생 데이터의 예측 결과 $Y'_{f\to1}$에서 환자들의 사망 경향이 평균적으로 증가할 것이므로 평균값 $E(Y'_{f\to1})$가 증가한다. 또, 미발생 시뮬레이션 데이터의 예측 결과 $Y'_{f\to0}$에서 환자들의 사망 경향이 평균적으로 감소할 것이므로 평균값 $E(Y'_{f\to0})$는 감소하게 된다. 따라서, 의료 이벤트 f가 사망 마커일 수록 $E(Y'_{f\to1}) - E(Y'_{f\to0})$는 큰 양수 값이 된다.

의료 이벤트 f가 생존 마커 이벤트이면 (즉, 사망과 음의 상관관계를 가지는 이벤트이면), 사망 마커 점수 $IS(f)$는 작은 음수 값을 가진다.

의료 이벤트 f가 보편적으로 발생하는 이벤트 일수록 $IS(f)$의 절대값은 커진다 (즉, 보편적으로 발생하는 이벤트에 높은 중요도 점수를 부여한다). 왜냐하면 수식에서 $Entropy(X_f)$는 의료 이벤트 f가 보편적으로 발생하면 커지고, 희귀하게 발생하면 작아지는 양수 값이기 때문이다.

4. 실험 및 결과

4.1 데이터 처리

우리는 MIMIC-III 데이터베이스에서, 의료 이벤트 종류로 구성된 26개의 CSV파일을 다운로드 받고 데이터 처리를 수행하였다. 환자 명수 N=7,727, 시점 개수 T=10, 의료이벤트 개수 F=4,068인 3차원 이진 설명변수 데이터 X와, 각 폐렴 환자가 사망 or 퇴원하였는지에 따라 1 or 0으로 이진 라벨링된 목적변수 데이터 Y를 생성하였다.

4.2 시계열 예측 모델의 정확도

우리는 폐렴 환자의 사망/생존 예측에 대해서, 학습:시험=8:2 비율의 10겹 교차검증 방법과 ROC-AUC 정확도를 사용하여, LSTM 모델의 예측 정확도를 비시계열 모델(의사결정나무, KNN, 베르누이 나이브베이즈, MLP, AdaBoost, 랜덤포레스트, GradientBoost) 및 다른 시계열 모델(RNN, GRU)과 비교하였다[그림 3]. 비교 결과, 시계열 모델이 비시계열 모델보다 ROC-AUC 정확도가 더 높았으며, 시계열 중에서도 LSTM 모델이 가장 높은 ROC-AUC 정확도(0.74)를 보여주었다.

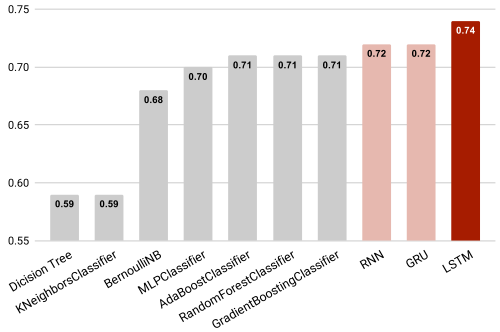

그림 3 폐렴 환자 사망 예측에 대한 ROC-AUC 비교. 비시계열 모델(회색) vs 시계열 모델 (연빨강) vs LSTM (진빨강).

4.3 폐렴 환자의 사망/생존 의료 이벤트

제안된 방법으로 총 4068개 의료 이벤트의 사망 마커 점수에 대해서, 양의 절댓값 top 4 이벤트와 음의 절댓값 top 4 이벤트를 각각 사망/생존 마커 이벤트로 선별하였다. 문헌 조사 결과, 사망/생존 마커에 대해서 폐렴의 악화/완화와 관련된 문헌을 찾을 수 있었다. [표1, 2]. 이벤트 발생 비율을 조사하였을 때, 사망 마커는 사망한 환자에서 더 높은 빈도로 발생하고[그림 4A], 생존 마커는 생존한 환자에서 더 높은 빈도로 발생하는 것을 확인할 수 있었다[그림 4B].

표 1 사망 마커 TOP 4에 대한 폐렴 관련 문헌

이벤트	점수	폐렴 관련 문헌
적혈구 용적 분포폭치 (RDW)	0.108	RDW 수치가 정상범위인 13.8%보다 높으면 사망률이 약 1.7배 상승한다는 연구결과가 있다. [5]
혈액요소질소 (Blood Urea Nitrogen)	0.085	신장기능이상을 측정하는 검사 항목. 신장이 건강하지 않은 폐렴 환자가 사망할 위험이 높다는 연구결과가 있다. [6]
크레아티닌 (Creatinine)	0.058	폐렴이 악화되었을 때, 그 사망군에서 비정상적인 크레아티닌 수치가 검출된 경우가 발견됐다는 연구결과가 있다. [7]
황산 모르핀 (Morphine Sulfate)	0.045	중환자실에서 가장 많이 사용되는 마약성 진통제. 중증 환자에 대한 진통제로 황산 모르핀이 추천되기도 한다. [8]

표 2 생존 마커 TOP 4에 대한 폐렴 관련 문헌

이벤트	점수	폐렴 관련 문헌
헤파린 (Heparin)	-0.054	급성 호흡 부전을 동반한 폐렴 환자에게 실시하는 저분자 헤파린요법은 사망률을 줄인다는 보고가 있다. [9]
알라닌 전달 효소 (ALT)	-0.040	특별한 관련 문헌을 찾을 수 없음.
판토프라졸 (Pantoprazole)	-0.037	특별한 관련 문헌을 찾을 수 없음.
포도당 (Dextrose)	-0.033	정상 범위 내에서 포도당 수치를 감시하고 유지하는 것은 성공적인 치료의 중요한 요소다. [10]

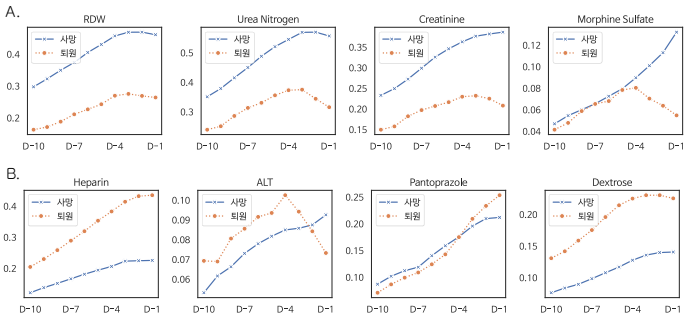

그림 4 사망 및 생존 환자에 대한 이벤트 발생 빈도 비율. (A) 사망 마커 (B) 생존 마커. 가로축은 사망/퇴원일(D0) 전 10일(D-10 ~ D-1)을, 세로축은 이벤트 발생 빈도 비율을 나타낸다.

5. 결 론

본 연구에서는 중환자실 폐렴 환자의 전자 의무 기록을 기반으로 생존 예측 모델을 구축하고 시뮬레이션에 기반한 시계열 사망 마커 탐지 기법을 제안하였다. MIMIC-III 데이터에서 폐렴 환자의 사망/생존 마커를 탐지하였을 때, 사망/생존에 관련된 의료 이벤트를 검출할 수 있음을 확인하였다.

6. 참고문헌

[1] Johnson AE, et al. MIMIC-III, a freely accessible critical care database. Scientific data, 3(1):1-9. 2016.
[2] Abebe TG, et al. Model-free feature selection to facilitate automatic discovery of divergent subgroups in tabular data. arXiv e-prints. 2022.
[3] Scheurwegs E, et al. Selecting relevant features from the electronic health record for clinical code prediction. Journal of biomedical informatics, 74:92-103. 2017.
[4] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation, 9(8):1735-1780. 1997.
[5] Yoo KD, et al. Red blood cell distribution width as a predictor of mortality among patients regularly visiting the nephrology outpatient clinic. Scientific reports, 11(1):1-9. 2021.
[6] Feng DY, et al. Elevated blood urea nitrogen-to-serum albumin ratio as a factor that negatively affects the mortality of patients with hospital-acquired pneumonia. Canadian Journal of Infectious Diseases and Medical Microbiology. 2019.
[7] 장상민, et al. 응급실을 통해 중환자실에 입원한 패혈증 환자의 예후 예측. 대한응급의학회지, 27(4):306-312. 2016.
[8] Shapiro BA, et al. Practice parameters for systemic intravenous analgesia and sedation for adult patients in the intensive care unit: an executive summary. Critical care medicine, 23:1596-1600. 1995.
[9] 이미숙, et al. 성인 지역사회획득 폐렴 항생제 사용지침. Infection & Chemotherapy. 2018.
[10] Kubisz A, et al. Elevated blood glucose level as a risk factor of hospital-acquired pneumonia among patients treated in the intensiv care unit (ICU). Przeglad Lekarski, 68(3):136-9. 2011.