DEC기반의 비소세포성 폐암환자 클러스터링*

최철웅[0] 염성웅 김경백
전남대학교 인공지능융합학과
sentilemon02@gmail.com, yeomsw0421@gmail.com, kyungbaekkim@jnu.ac.kr

DEC-based Non-small cell lung cancer patient clustering

Chulwoong Choi[0], Sungwoong Yeom, Kyungbaek Kim
Department of Artificial Intelligence Convergence, Chonnam National University

요 약

병원에서는 미국 암 연합회(AJCC)의 TNM병기 분류체계를 바탕으로 폐암환자의 치료방침 및 예후를 결정하고 있다. 하지만 암세포의 크기, 림프절 전이, 기타장기 전이 정보를 기준으로 분류되는 TNM병기 분류체계는 환자의 다양한 특성을 반영하기 어렵다. 폐암환자의 임상정보와 영상정보를 함께 사용하고 임베딩을 통한 차원축소 과정을 진행하면 환자의 다양한 특성을 반영하여 군집을 나눌 수 있다. 이 논문에서는 폐암환자의 다중 모드 데이터(CLINICAL, PET)에서 특징을 추출하고 추출 된 특징과 DEC(Deep Embedding for Clustering) 모델을 사용하여 폐암환자 군집분석을 진행하였다. TNM병기 분류체계, K-Means와 DEC 군집분석 결과를 Cox비례위험모형, 카플란마이어 생존곡선, Boxplot 그래프를 사용하여 비교분석하였다. K를 4로 설정하고 DEC 군집분석을 사용하였을 때 TNM병기 분류체계와 K-means를 사용하였을 때 보다 생존군집의 응집도가 높고 생존시간예측 정확도가 높은 것을 확인하였다.

1. 서 론

폐암환자들의 치료 및 예후를 결정하는 TNM병기 분류체계는 지속적인 수정을 통해 개정되고 있지만 병기분류에 사용되는 환자의 특성(T, N, M)이 정해져 있기 때문에 생존분석과 같이 환자 개개인의 다양한 특성을 반영해야하는 경우 한계가 있다[1]. 최철웅, 김경백[2]의 연구에서는 폐암환자의 TNM병기만을 사용하고 다양한 클러스터링 기법을 통해 최종병기를 결정할 수 있는지 알아보았다. 데이터과학분야에서 확인했을 때 미국 암 연합회의 TNM병기 분류체계는 한계가 있었고 생존과 연관이 높은 새로운 최종병기 군집이 필요함을 알 수 있었다.

폐암 환자의 다양한 특성을 반영하여 생존군집을 나누는 클러스터링 모델을 만들기 위해서는 임상(Clinical)정보와 PET 영상을 함께 사용하는 다중 모드 데이터를 사용해야 한다[3].

이 논문에서는 폐암환자의 다중 모드 데이터를 사용하여 특징을 추출하고 딥러닝기반의 DEC(Deep Embedding for Clustering) 기법을 통해 클러스터링 한다. Cox비례위험모형, 카플란마이어 생존 곡선과 Boxplot을 사용하여 TNM병기분류와 클러스터링 결과를 비교분석한다.

2장에서는 특징 추출과 DEC기법에 대해 알아보고, 3장에서는 다중 모드 데이터와 DEC를 사용한 생존 군집분석에 대해 설명한다. 4장에서는 실험을 통해 TNM병기분류와 다중 모드 데이터를 사용한 DEC기법의 성능을 비교하고, 5장에서는 결론 및 향후 연구에 대해 기술한다.

* 이 논문은 2019년도 정부(과기정통부)의 재원으로 한국연구재단 바이오·의료기술개발사업의 지원을 받아 수행된 연구임(NRF-2019M3E5D1A02067961).

2. 관련연구

2.1 특징 추출(Feature Extraction)

폐암환자의 종양 크기와 위치를 확인할 때 사용하는 PET(Positron Emission Tomography) 영상은 128x128픽셀의 2차원 데이터로 한 사람의 전신을 표현하기위해서는 신장에 따라 약 250~450장의 2차원 데이터 구성되며 의학영상 정보의 국제 표준인 DICOM(Digital Imaging and Communications in Medicine) 포맷으로 저장된다.

PET 영상은 환자의 신장에 따라 구성되는 이미지 수가 상이하기 때문에 클러스터링을 위해서는 PET영상을 일정한 사이즈로 정규화 하는 과정이 필요하다. 보간법을 사용한 전처리 과정을 통해 약 250~450장의 2차원 데이터를 50장으로 압축하였다. 최종적으로 폐암 환자 1명당 50x128x128 사이즈의 PET 3D 이미지를 생성하였으며, 3차원 PET 이미지의 해상도는 약 82만 픽셀이 된다.

이러한 대용량 이미지 정보를 사용하여 클러스터링을 진행하기 위해서는, 전체 이미지에서 특징 벡터를 추출하는 과정이 필요하다. 최근, CNN(Convolutional Neural Network) 딥러닝 모델을 바탕으로 대용량 이미지 정보에서 특징을 추출하는 연구가 활발히 진행 중이다[4][5][6]. [4][5][6]의 연구에서는 CNN기반의 DenseNet, AlexNet 및 VGG16 모델 등을 사용하여 각각 흉부 X-ray영상, 손 제스처 영상 및 숫자필기 이미지에서 특징을 추출하는 연구가 진행되었다.

이 연구에서는 DICOM포맷의료영상 특징추출에 가장 많이 사용하는 3D RESNET-18 모델을 사용하여 특징을 추출하였다. 3D RESNET 모델은 RESNET 모델에 3D 영

상을 사용할 수 있게 개량한 모델이다. 50x128x128 크기의 PET 3D 영상을 입력으로 사용하여 최종적으로 환자 1명당 512개의 특징 벡터를 추출하였다.

2.2 DEC(Deep Embedding for Clustering)

J.Xie, R.Girshick, A.Farhadi[7]의 DEC 모델은 기존에 클러스터링 모델로 많이 사용하는 K-Means와 DBSCAN 등의 머신러닝 모델과 다르게 인공지능을 활용한 확률기반 클러스터링 모델이다. DEC 모델은 “curse of dimensionality(차원의 저주)”를 피하기 위해 오토인코더를 사용하여 입력값을 Feature Representation(특징표현)한다. 재표현된 특징을 바탕으로 t-분포를 사용하여 특징들을 소프트 할당(Soft Assignment)한다. 마지막으로 쿨백-라이블러 발산(Kullback-Leibler divergence)을 사용하여 두 확률분포의 차이를 계산하고 차이가 가장 작아지도록 오차역전파(Back Propagation)를 통해 모델을 학습한다. 기존 머신러닝 모델과의 가장 큰 차이점은 타켓분포를 사용하여 비지도학습인 클러스터링을 지도학습처럼 사용한 점이다.

이 연구에서는 3D RESNET-18로 추출된 512개 특징벡터를 DEC 기법을 사용하여 클러스터링한다.

3. 다중모드 데이터와 DEC 기반의 폐암환자 클러스터링

이 논문에서는 폐암환자의 다양한 특성을 반영한 생존군집 클러스터링을 위해 다중 모드 데이터를 사용한 DEC기반의 폐암환자 클러스터링 방법을 제안한다. 제안한 모델의 진행과정은 그림 1과 같다. 환자의 다양한 특성을 반영하기 위해서는 다중 모드 데이터를 사용해야한다. 하지만, 다중 모드 데이터를 사용하게 되면 환자의 특성이 매우 많아지기 때문에 모델이 학습하는 도중 ‘차원의 저주’에 빠질 확률이 높다. 따라서, 임베딩을 통한 차원축소 및 특징표현 등의 과정이 필요하다.

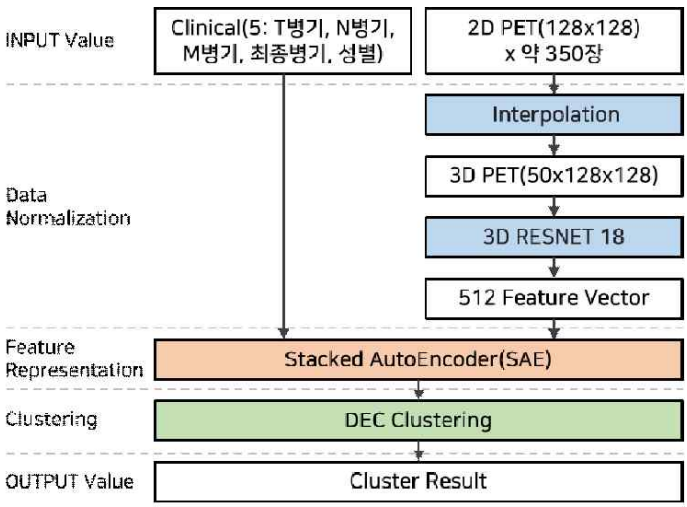

그림 1. 다중모드 데이터와 DEC 기반 폐암환자 클러스터링 과정

폐암환자의 임상데이터(Clinical)와 영상데이터(PET)를 입력으로 사용하며 표1과 같다. 임상데이터 경우 폐암과 관련이 높은 5개의 특징만을 사용하였다.

표 1. 데이터셋 설명

데이터 타입	내용
Clinical(5)	최종병기, 암 세포의 크기(T병기), 림프절 전이 정도(N병기), 타 장기 전이 여부(M병기), 성별
2D PET (128x128)	폐암환자의 PET이미지, 성인기준 신장에 따라 약 250~450장으로 구성

2D PET영상은 정규화를 위해 보간법(Interpolation)을 사용하여 50x128x128 사이즈의 3D PET영상을 생성한다.

3D RESNET-18 모델을 사용하여 3D PET영상에서 512개의 특징벡터를 추출한다.

5개의 임상정보와 512개의 3D PET 특징벡터를 포함한 총 517개의 특징벡터를 Stacked AutoEncoder(SAE)를 사용한 Feature Representation을 통해 10개의 특징벡터로 재 표현한다.

재 표현된 특징벡터를 바탕으로 DEC 클러스터링을 진행한다.

4. 실험 및 분석

실험에 사용한 데이터셋은 화순전남대학교병원의 비소세포폐암(NSCLC) 환자 2687명의 임상정보와 영상정보를 활용하였다. 임상정보와 영상정보는 표1에서 설명한 5개의 임상정보와 PET 영상정보를 사용하였다.

클러스터링 결과와 최종병기(1기, 2기, 3기, 4기)를 비교하기 위해 K-value는 4를 사용한다.

클러스터링 결과를 비교분석하기위해 표2의 3개의 모델을 사용하였다. 최종병기 모델은 병원에서 사용하는 미국암연합회(AJCC)의 TNM 병기분류체계를 바탕으로 결정되는 최종병기(1기, 2기, 3기, 4기)를 사용하였다. K-Means 모델은 그림1의 Data Normalization과정까지 동일하게 진행 후 K-Means 알고리즘을 사용하여 클러스터링을 진행하는 모델이다. n_init=20, max_iter=300으로 설정하여 K-Means 알고리즘이 다른 중심 시드로 실행되는 횟수는 20회, 단일 실행에 대한 K-Means 알고리즘의 최대 반복 횟수는 300회로 설정하였다.

표 2. 실험에 사용한 모델

모델	설명
최종병기 모델	폐암환자의 임상정보인 최종병기를 사용한 클러스터링 모델
K-Means 모델	임상정보와 PET정보를 함께 사용한 K-Means 모델
DEC 모델	이 논문에서 제안한 임상정보와 PET정보를 함께 사용한 DEC 모델

클러스터링 모델의 평가지표로는 Cox비례위험모형, 카플란마이어 생존곡선, 박스플롯(Boxplot)그래프를 사용하였다.

표 3은 클러스터링 모델별 Cox비례위험모형을 사용하여 C-index평가결과를 나타내고 있다. C-index는 사망여부를 고려한 생존시간 예측정확도를 계산할 때 가장 많이 사용하는 지표로 이 논문에서 제안한 DEC모델이 0.73으로 가장 좋은 성능을 보이고 있다.

표 3. 클러스터링 모델별
Cox비례위험모형의 C-index 평가결과

모델	C-index
최종병기 모델	0.6960
K-Means 모델	0.6591
DEC 모델	0.73

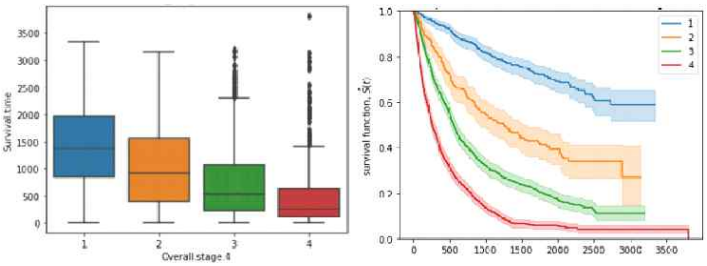

그림 2. 최종병기 모델의 카플란마이어 생존곡선과 Boxplot그래프 결과

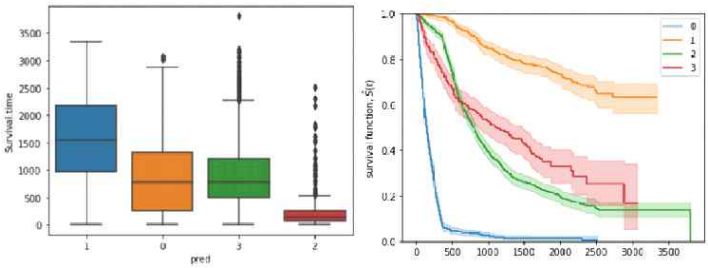

그림 3. K-Means 모델의 카플란마이어 생존곡선과 Boxplot그래프 결과

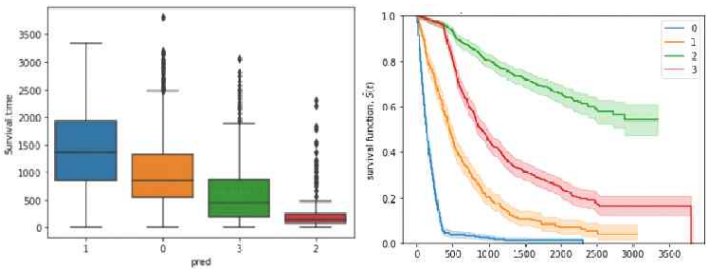

그림 4. DEC 모델의 카플란마이어 생존곡선과 Boxplot그래프 결과

클러스터링 모델별 박스플롯 그래프(좌)와 카플란마이어 생존곡선(우)은 그림2, 그림3, 그림4를 통해 확인할 수 있다.

카플란마이어 생존곡선은 관찰시간에 따라서 사건 발생시점의 사건 발생율을 계산하는 방법이다. K-Means의 경우 2번 군집과 3번 군집의 생존곡선이 겹침으로써 성능이 좋지 않은 것을 확인할 수 있다. 반면 최종병기 모델과 DEC 모델 모두 4개의 군집이 서로 겹치지 않게 그려졌으며 DEC모델의 생존곡선이 최종병기모델보다 더 확장 된 스펙트럼을 보이며 폭넓은 생존군집 분포를 확인할 수 있다.

박스플롯그래프는 데이터의 대체적인 분포 형태를 쉽게 확인할 수 있다. 최종병기 모델과 K-Means 모델의 박스플롯그래프는 4개의 군집에 포함된 환자의 생존일이 대다수 중복되는 것을 확인할 수 있다. 반면에 DEC 모델의 박사플롯그래프는 다른 모델에 비해 환자의 생존일이 소량 중복되며 환자의 생존일수에 맞게 클러스터링 된 것을 확인할 수 있다.

5. 결론 및 향후 연구

이 논문에서는 최종병기가 아닌 새로운 폐암환자의 생존군집을 찾아내기 위해 폐암환자의 다중 모드 데이터를 사용하여 특징을 추출하고 딥러닝기반의 DEC(Deep Embedding for Clustering) 기법을 통해 폐암환자를 클러스터링하고 비교분석하였다. DEC모델이 최종병기모델과 K-Means 모델에 비해 모든 지표에서 우수한 성능을 보였다. 하지만 DEC모델의 성능을 최종병기모델과 비교하기 위해 K-value를 4로 고정하고 모델을 학습했기 때문에 현재의 모델이 가장 좋은 성능을 나타낸다고 볼 수 없다. 따라서, DEC 모델 최적의 K-value를 찾는 연구가 추가적으로 진행되어야 한다.

참고문헌

[1] 김혜영, “폐암의 병기 결정”, 대한의사협회 대한의사협회지, 제51권, 제12호, 1118-1124쪽, 2008년

[2] 최철웅, 김경백, “폐암환자 생존분석에 대한 TNM 병기 군집분석 평가”, 한국스마트미디어학회 스마트미디어저널, 제9권, 제4호, 126-133쪽, 2020년

[3] 최철웅, 김현지, 김경백 외 4명, “다중 모드 데이터를 사용한 폐암 생존분석 검토”, 한국정보처리학회 추계학술발표대회 논문집, 제27권, 제2호, 784-787쪽, 2020년

[4] Varshni, Dimpy, et al. “Pneumonia detection using CNN based feature extraction.“ 2019 IEEE International Conference on Electrical, Computer and Communication Technologies(ICECCT), 2019.

[5] Barbhuiya, Abul Abbas, Ram Kumar Karsh, and Rahul Jain. “CNN based feature extraction and classification for sign language.“ Multimedia Tools and Applications Vol. 80, No. 2, pp. 3051-3069, 2021.

[6] Zhao, Hui-huang, and Han Liu. “Multiple classifiers fusion and CNN feature extraction for handwritten digits recognition.“ Granular Computing, Vol. 5, No. 3, pp. 411-418, 2020.

[7] Xie, Junyuan, Ross Girshick, and Ali Farhadi. “Unsupervised deep embedding for clustering analysis.“ International conference on machine learning, PMLR Vol. 48, pp. 478-487, 2016.