

폐암환자 생존분석에 대한 TNM 병기 군집분석 평가

(Accessing the Clustering of TNM Stages on Survival Analysis of Lung Cancer Patient)

최철웅*, 김경백**

(Chulwoong Choi, Kyungbaek Kim)

요약

병원에서는 폐암 환자의 최종병기를 기준으로 치료방침 및 예후를 결정하고 있다. 폐암 환자의 최종병기는 미국 암 연합회(AJCC)에서 제공하는 TNM 분류방법을 바탕으로 7단계로 나누어 진단된다. 이런 접근 방법은 환자의 치료, 예후 및 생존일 예측 등 다양한 분야에서 사용하기에 한계가 있다. 이 논문에서는 데이터 과학적 접근을 통해 T, N, M병기를 사용하여 생존일 수별 환자집단을 나눌 수 있는지 알아보기 위해 비지도 학습 중 하나인 군집분석(Clustering)을 진행한 후 군집분석의 결과를 Cox비례위험모형을 사용하여 비교 하였다. 환자들의 최종병기를 사용하지 않고, T, N, M병기 정보만 사용하였을 때 생존시간 예측정확도가 더 높은 것을 확인하였다. 특히, AJCC의 최종병기 7단계와 같이 군집의 개수를 7로 설정했을 때보다 군집의 수를 축소하거나 확장했을 때 T, N, M 병기 군집분석을 통한 생존시간 예측정확도가 향상하는 것을 확인하였다.

■ 중심어 : 폐암 ; 클러스터링 ; TNM병기 ; Cox비례위험모형 ; 카플란-마이어 생존 곡선

Abstract

The treatment policy and prognosis are determined based on the final stage of lung cancer patients. The final stage of lung cancer patients is determined based on the T, N, and M stage classification table provided by the American Cancer Society (AJCC). However, the final stage of AJCC has limitations in its use for various fields such as patient treatment, prognosis and survival days prediction. In this paper, clustering algorithm which is one of non-supervised learning algorithms was assessed in order to check whether using only T, N, M stages with a data science method is effective for classifying the group of patients in the aspect of survival days. The final stage groups and T, N, M stage clustering groups of lung cancer patients were compared by using the cox proportional hazard model. It is confirmed that the accuracy of prediction of survival days with only T, N, M stages becomes higher than the accuracy with the final stages of patients. Especially, the accuracy of prediction of survival days with clustering of T, N, M stages improves when more or less clusters are analyzed than the seven clusters which is same to the number of final stage of AJCC.

■ keywords : Lung Cancer ; Clustering ; TNM Stage ; Survival Analysis ; Kaplan-Meier Survival Curve

I. 서론

국내 암 사망률 1위인 폐암은 최종병기가 1기에서 4기로 진행되는 데 1년이 걸리지 않을 정도로 매우 빠르게 확산된다. 최종병기를 정확하게 판단해야만 올바른 치료방침과 예후를 결정할 수 있다[1]. 현재 최종병기는 종양의 크기에 따른 T병기, 림프절 전이 정도에 따른 N병기, 다른 장기의 원격전이에 따른 M병기 3가지 병기를 사용한 TNM 분류방법을 기반으로 한다[2].

국제적으로 가장 많이 사용하는 TNM 분류방법은 미국암연합회(American Joint Committee on Cancer, AJCC)에서 작성한 암 병기 매뉴얼이다. TNM 분류방법이 암별로 상이하지만 폐암 또한 AJCC의 T, N, M병기 분류표를 기준으로 최종병기를 결정한다. T, N, M병기 분류표가 포함되어 있는 암 병기 매뉴얼은 2017년 8판으로 개정되었지만 현재 병원에서는 7판을 주로 사용하고 있다. TNM병기를 기준으로 만들어진 AJCC의 폐암환자 T, N, M병기 분류는 표 1과 같다[3].

AJCC의 암 병기 매뉴얼이 개정되면서 폐암 병기를 검증하려

* 정회원, 전남대학교 인공지능융합학과 박사과정

** 정회원, 전남대학교 인공지능융합학과 교수

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017R1A2B4012559). 이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단 바이오·의료기술개발사업의 지원을 받아 수행된 연구임(NRF-2019M3E5D1A02067961).

접수일자 : 2020년 08월 18일

수정일자 : 2020년 11월 06일

게재확정일 : 2020년 11월 10일

교신저자 : 김경백 e-mail : kyungbaekkim@jnu.ac.kr

는 많은 연구들이 진행되었다. [4-9]의 연구들은 TNM병기를 사용하여 개정된 폐암 병기의 타당성을 검증하였다. 환자의 치료, 예후, 생존 등과 같이 다양한 상황을 고려하며 병기를 검증하였지만 대부분의 연구가 일반적인 통계방법만을 사용하여 검증되었다는 한계가 있다.

최철웅, 벨루루간 아레수 발라지, 김경백[10]의 연구에서는 데이터과학 분야에서 폐암 환자의 TNM병기만을 사용하여 최종병기를 결정할 수 있는지 알아 보기위해 비지도학습인 K-Means클러스터링을 사용하였다. 환자들이 AJCC의 최종병기에 맞게 군집되지 않았으며 정제되지 않은 데이터일수록 성능은 하향됐다. 데이터과학분야에서 확인했을 때 현재 병원에서 사용하는 AJCC의 최종병기 7단계는 한계가 있으며 수정이 필요하다.

표 1. AJCC의 폐암 TNM병기 분류방법(7판)과 최종병기

최종병기	T, N, M 병기
1A	T1, N0, M0
1B	T2, N0, M0
2A	T1, N1, M0
2B	T2, N1, M0
	T3, N0, M0
3A	T3, N1, M0
	T1, N2, M0
	T2, N2, M0
	T3, N2, M0
3B	T4, N0, M0
	T4, N1, M0
	T4, N2, M0
	T1, N3, M0
	T3, N3, M0
	T4, N3, M0
4	Any T, Any N, M1

논문에서는 데이터 과학적 접근법을 통해 T, N, M 병기정보를 이용해 폐암환자 생존분석의 가능성을 확인하고자 한다. 이를 위해, 다양한 클러스터링 알고리즘을 사용하여 환자의 T, N, M 3가지 병기를 클러스터링한다. 클러스터링 결과와 AJCC의 7단계 최종병기를 카플란마이어 생존 곡선과 Cox비례위험모형을 사용하여 폐암환자 생존시간에 미치는 영향 정도를 평가한다. 화순전남대학교병원에서 수집된 폐암환자의 생존일수와 T, N, M 병기 및 최종병기 기록을 사용하여 생존분석을 수행

하였다.

이 논문은 다음과 같이 구성된다. 2장에서는 군집분석, 중도절단 자료, 카플란 마이어 생존 곡선 및 Cox비례위험모형에 대해 알아보고, 3장에서는 실험에 사용한 데이터셋에 대해 설명한다. 4장에서는 다양한 군집분석 알고리즘을 사용하여 TNM병기 클러스터링을 진행하고, 5장에서는 군집분석 결과를 바탕으로 Cox비례위험모형을 사용하여 비교한다. 6장에서는 비교결과를 토대로 결론 및 향후연구에 대해 기술한다.

II. 관련 연구

1. 군집분석(Clustering)

군집분석(Clustering) 또는 클러스터링은 비지도학습 방법 중 한가지이다. 클러스터링을 사용하면 유사한 개체들을 그룹으로 나눌 수 있어 토픽 분류[18], 무선 센서 네트워크[19] 등 다양한 분야에서 활용되고 있다. 대표적인 클러스터링 방법으로는 K-Means 클러스터링, DBSCAN 클러스터링, Mean-Shift 클러스터링 및 Agglomerative 클러스터링 그리고 방법 등이 있다.

K-Means 클러스터링은 학습 전 군집해야 할 클러스터 수 K를 알려주고 학습을 진행한다. K개의 무작위 중심점(센트로이드)을 찍은 후, 유클리드 거리를 통해 중심점에서 각 점간의 거리 합을 계산하며 중심점을 이동시킨다. 각 점간의 거리 합이 가장 작을 때 해당 중심점을 기준으로 클러스터링 하는 방법이다. 중심점은 각 클러스터에 속한 데이터들의 평균값을 위치로 갖는다[11,17]. 그림 1은 K를 3으로 설정하고 진행한 K-Means 클러스터링의 예시이다. 빨간점은 클러스터의 센트로이드로 중심점을 의미한다.

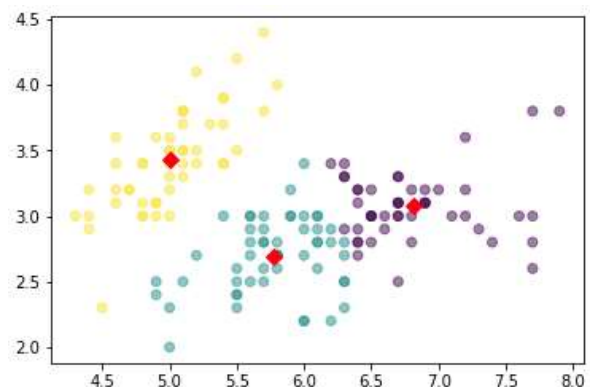


그림 1. K-Means 클러스터링 시각화

DBSCAN 클러스터링은 밀도 기반 알고리즘으로 군집해야 할 클러스터 수를 지정할 필요가 없으며 무작위로 고른 시작점

을 기준으로 이웃 안에 충분한 수의 점이 들어있다면 군집한다.

Agglomerative 클러스터링은 병합군집 기반 알고리즘으로 시작할 때 각 포인트를 하나의 클러스터로 지정하고 종료 조건을 만족할 때까지 가장 비슷한 클러스터를 병합한다.

Mean-Shift 클러스터링은 슬라이딩 윈도우(Sliding-Window)에 기반을 둔 방법으로 슬라이딩 윈도우 내에 있는 점들의 평균이 되고자 하는 중심점의 후보를 기준으로 군집한다 [12].

2. 중도절단 자료

폐암 생존분석시 가장 이상적인 데이터는 모든 참여자(폐암 환자)가 사망 할 때까지 관찰 된 데이터다. 그러나 사망하기 전 관찰기간이 종료될 수도 있고 관찰기간이 종료되기 전 사망하는 경우도 있기 때문에 대부분 불가능하다. 관찰이 끝나기 전 사망여부를 확인할 수 없는 자료를 중도절단 자료(Censored Data)라고 한다. 생존분석에서는 중도절단 자료를 명확하게 구분해야만 편향(bias)을 줄이고 성능을 향상시킬 수 있다[13].

3. 카플란마이어 생존 곡선

카플란 마이어 생존 곡선(Kaplan-Meier Survival Curve)은 시간별로 환자가 사망하는 이벤트 발생시점마다 구간생존율을 계산하여 그리는 계단형태의 생존곡선이다[13]. 카플란 마이어 생존 곡선을 그리기 위해서는 생존여부, 생존기간 그리고 요인 3가지를 사용하여 생존율을 계산한다. 시간은 x축에 표시되고 각 시점에서 생존자의 비율은 y축에 표시된다.

4. Cox비례위험모형

Cox비례위험모형(Cox Proportional Hazard Model)은 준모수방법으로 특정 변수에 대한 relative hazard를 계산할 수 있고 특정분포를 가정하지 않기 때문에 많이 사용하는 생존분석 모델이다.

$$h(t | X_i) = h_0(t) \exp(X_i \cdot \beta) \quad (1)$$

시간과 이벤트 데이터를 사용하여 특정시간 t에 중요 이벤트가 발생할 확률 relative hazard는 수식(1)을 통해 계산할 수 있다. h_0 은 시간에 따라 변화하는 위험함수이고, X_i 는 대상 i 의 변수, β 는 모델이 학습해야 할 계수를 뜻한다[14].

생성된 Cox 모델의 생존시간 예측정확도는 c-index를 통해 측정할 수 있다. c-index는 수식(2)와 같이 계산할 수 있다[15]. c-index의 경우 0.0~1.0사이의 값을 가지며 1.0에 가까울수록

정확하며, 0.5에 가까울수록 무작위로 예측된다고 해석한다.

$$c = \frac{\sum_{i \in U} \{ \sum_{T_j > T_i} 1_{f_j > f_i} \}}{\sum_{i \in U} \{ \sum_{T_j > T_i} 1 \}} \quad (2)$$

III. 실험 데이터셋

1. 실험 데이터셋

화순전남대학교병원에서 제공받은 폐암환자 데이터셋을 사용하였으며 표 2와 같다. 환자 수는 4400명이며 14개의 컬럼으로 구성된다. 본 논문에서 T, N, M 병기와 최종병기(4, 7), 생존일수, 생존이벤트 7개의 컬럼을 사용한다.

표 2. 폐암환자 데이터셋

환자 수	4400명	
컬럼 수	14개	
T 병기(Clinical.T.Stage)	1, 2, 3, 4	
N 병기(Clinical.N.stage)	0, 1, 2, 3	
M 병기(Clinical.M.stage)	0, 1	
최종병기 (Final Stage)	4(Overall.stage_4)	1, 2, 3, 4
	7(Overall.stage)	1A, 1B, 2A, 2B, 3A, 3B, 4
생존일수(Survival.time)	2일~5554일	
생존이벤트(DeadStatus.event)	0, 1, 9	

표 2의 첫 번째 열의 괄호안의 값은 실제 데이터셋에서 사용하는 컬럼명이다. T병기는 종양의 크기를 나타내며 T1~T4까지 4개로 구분되며 숫자가 커질수록 종양의 크기도 커진다. N 병기는 림프절의 전이정도를 나타내며 N0~N3까지 4개로 구분되며 숫자가 커질수록 림프절의 전이개수가 많아진다. M병기는 원격전이를 나타내며 다른 장기로 전이가 되지 않은 경우 M0, 전이가 된 경우 M1이다. 최종병기는 1기, 2기, 3기, 4기 4개로 구분하는 최종병기4와 1A기, 1B기, 2A기, 2B기, 3A기, 3B기, 4기 7개로 구분하는 최종병기7이 있다. 생존일수는 폐암환자들이 생존한 일수로 2일부터 5554일 사이의 값을 갖는다. 마지막으로 생존이벤트는 환자의 생존여부를 나타내며 0은 생존, 1은 사망, 9는 치료 도중 환자의 의사에 따라 치료를 중단한 환자를 뜻한다.

최종병기별 폐암환자 수는 표3과 같다. 각 비율은 1기(19.16%), 2기(9.5%), 3기(34.84%), 4기(36.5%)와 같다. 병의 진행속도가 빠른 폐암의 특성상 1기, 2기 환자보다는 3기, 4기에 환자가 대부분 밀집되어 있으며 약 71.3%비율을 차지하고 있다.

2. 데이터 전처리

T병기는 1~4, N병기는 0~3, M병기는 0~1 사이의 값으로 구성되며 값의 범위가 서로 다르다. Scikit-Learn에서 제공하는 MinMaxScaler() 함수를 사용하여 T, N, M병기를 0~1 사이의 값으로 정규화 한다.

생존이벤트 컬럼에서 관찰기간이 끝나기 전 치료를 중단한 환자(9)를 생존한 환자(0)로 수정하여 Censored Data를 0으로 취합한다.

최종병기를 7개로 분류하는 AJCC의 병기분류표를 재가공하여 1, 2, 3, 4기로 구분하는 새로운 최종병기4(Overall.stage_4)를 생성했다.

IV. TNM병기 클러스터링

폐암환자의 T, N, M병기만을 이용해서 클러스터링을 진행하였다. K값은 AJCC의 최종병기분류표와 같이 4와 7로 설정하였다. 4가지 성능평가 지표를 사용하여 클러스터링 성능을 비교 분석하였다.

1. 클러스터링 성능평가 척도

클러스터링 알고리즘은 비지도 학습방법으로 성능평가가 어렵기 때문에 다양한 성능평가 기준을 복합적으로 사용하여 성능을 평가한다. 표 4는 클러스터링 성능 평가시 보편적으로 많이 사용하는 ARI(Adjusted Rand Index), Homogeneity, Completeness와 V-Measure 성능평가 방법을 설명하고 있다. 모든 성능평가방법은 0.0부터 1.0사이의 값을 갖으며 1에 가까울수록 좋은 성능을 뜻한다[16].

2. 클러스터링 성능 비교분석

표 5는 폐암환자 4400명의 T, N, M병기만을 사용하여 K-Means와 Agglomerative로 클러스터링하고 성능을 비교한 결과다. K-Means와 Agglomerative 알고리즘 모두 K값이 7보다 4일 때 성능이 더욱 좋았다. K-Means(K=4)는 V-Measure가 0.79로 동일한 값들이 대부분 하나의 클러스터로 묶였고 이질적인 값이 적다. Agglomerative(K=4)도 V-Measure가 0.78로 성능이 K-Means(K=4)와 유사했다. 반면에 K-Means(K=7)와 Agglomerative(K=7) 모두 V-Measure는 0.63, ARI는 각각 0.38, 0.39로 AJCC의 최종병기 군집과 비교했을 때 매우 상이하게 군집됐다. 이외에도 K값을 설정하지 않는 DBSCAN과 Mean-Shift로 클러스터링한 결과 클러스터링 수는 각각 2개, 4개로 군집되었다.

표 3. 최종병기별 폐암 환자 수(4400명)

최종병기		환자 수	
1	1A	843명	420명
	1B		423명
2	2A	418명	238명
	2B		180명
3	3A	1533명	860명
	3B		673명
4	4	1606명	

표 4. 클러스터링 성능평가 방법

성능평가 방법	설명
ARI(Adjusted Rand Index)	두 클러스터간의 모든 쌍을 비교하며 유사성을 평가한다. Rand Index의 클러스터 수가 많아지면 점수가 높아지는 단점을 확률적으로 조정하여 보완한 값이다.
Homogeneity	각 클러스터마다 단일 클래스의 멤버인 데이터 포인트로만 모였는지를 평가한다.
Completeness	단일 클래스의 멤버인 모든 데이터 포인트가 하나의 클러스터로 모였는지를 평가한다.
V-Measure	Homogeneity와 Completeness가 얼마나 성공적으로 충족되었는지 명시적으로 측정하는 엔트로피 기반 평가 방법이다. Homogeneity와 Completeness의 조화평균을 계산한다.

표 5. 클러스터링 성능평가 방법

클러스터링	K-Means		Agglomerative	
	K=4	K=7	K=4	K=7
K Value	K=4	K=7	K=4	K=7
ARI	0.74	0.38	0.72	0.39
Homogeneity	0.80	0.66	0.80	0.66
Completeness	0.79	0.60	0.76	0.60
V-Measure	0.79	0.63	0.78	0.63

K-Means(K=4)의 ARI 0.74로 클러스터링 중 가장 좋은 성능을 보였지만 기존 AJCC의 병기 분류의 약 74%만 정확하게 군집됐기 때문에 데이터과학분야에서 T, N, M병기만을 가지고 AJCC의 최종병기를 분류하기는 어려워 보인다.

K-Means와 Agglomerative 알고리즘 모두 K값이 7보다 4일 때 성능이 더욱 좋았다. K-Means(K=4)는 V-Measure가 0.79로 동일한 값들이 대부분 하나의 클러스터로 묶였고 이질적인 값이 적다. Agglomerative(K=4)도 V-Measure가 0.78로 성능이 K-Means(K=4)와 유사했다. 반면에 K-Means(K=7)와 Agglomerative(K=7) 모두 V-Measure는 0.63, ARI는 각각 0.38, 0.39로 AJCC의 최종병기 군집과 비교했을 때 매우 상

이하게 군집됐다. 이외에도 K값을 설정하지 않는 DBSCAN과 Mean-Shift로 클러스터링한 결과 클러스터링 수는 각각 2개, 4개로 군집되었다. K-Means(K=4)의 ARI 0.74로 클러스터링 중 가장 좋은 성능을 보였지만 기존 AJCC의 병기 분류의 약 74%만 정확하게 군집됐기 때문에 데이터과학분야에서 T, N, M병기만을 가지고 AJCC의 최종병기를 분류하기는 어려워 보인다.

V. 폐암환자의 생존 분석

폐암환자의 TNM병기가 실제 생존시간에 미치는 영향 정도를 알아보기 위해 4장의 클러스터링 결과를 바탕으로 카플란마이어 생존 곡선과 Cox비례위험모형을 사용하여 생존분석하고 c-index를 통해 생존시간 예측정확도를 비교한다.

1. 상관분석

폐암환자의 생존일수와 다른 컬럼과의 상관관계를 알아보기 위해 14개의 컬럼 중 문자열 값으로 구성된 PatientID와 MCode를 제외한 12개의 컬럼을 사용하여 상관분석 하였다. 그림 2는 상관분석 결과를 히트맵 그래프로 시각화한 내용이다. 상관분석 결과는 -1~1사이의 값을 갖으며 1에 가까울수록 양의 상관관계(정비례), -1에 가까울수록 음의 상관관계(반비례), 0은 상관이 없음을 의미한다.

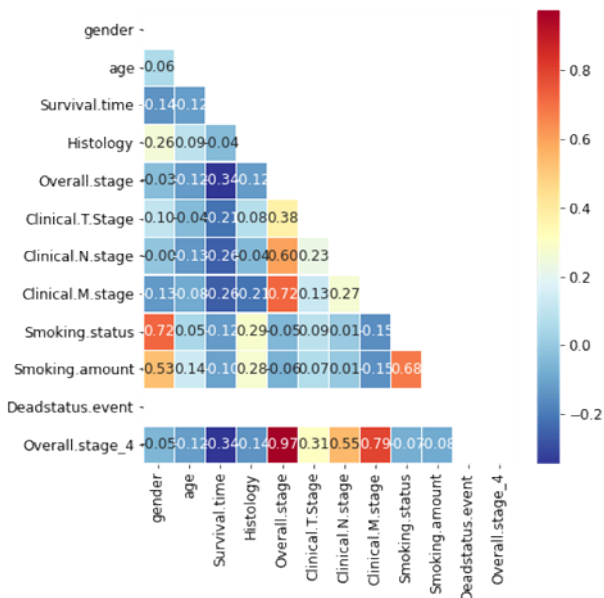


그림 2. 폐암환자 데이터 상관분석 히트맵

생존일(Survival.time)과 상관분석 결과 최종병기(7)과 최종 병기(4)가 0.34로 가장 상관이 높고 그 다음으로 M병기와 N병기가 0.26, T병기가 0.21 순으로 5개의 컬럼이 폐암환자의 생존

일과 연관이 높다.

2. 카플란마이어 생존 곡선

상관분석 결과를 바탕으로 상관이 높은 5개의 컬럼과 생존일 수, 생존이벤트를 사용하여 카플란마이어 생존 곡선을 그린다. 그림 3, 4, 5, 6은 최종병기, T병기, N병기, M병기를 기준으로 시각화한 카플란마이어 생존 곡선이다. 최종병기, T병기, N병기, M병기 모두 병기가 악화될수록 생존일수가 급격하게 감소한다. 상관분석 결과와 같이 카플란마이어 생존곡선 결과도 최종병기, T병기, N병기, M병기 모두 환자의 생존일과 연관이 많다. 그림 3의 폐암 3B, 4기 환자가 2년 이내에 빠르게 사망한다. 그림 4의 T3과 T4의 경우 뚜렷이 구분되는 T1, T2와 달리 곡선이 매우 유사한 경향을 띄며 폐암환자의 생존율이 매우 비슷하다. 그림 5의 N1, N2는 유사한 곡선을 띄며 폐암환자 생존율과 관련해서 뚜렷이 구분되지 않는다. 그림 6의 경우 M1은 그림 3의 폐암 4기 환자와 동일한 곡선이 그려진다. AJCC의 병기분류표에 의하면 M병기가 1이면 T병기와 N병기에 상관없이 모두 4기로 분류하기 때문이다.

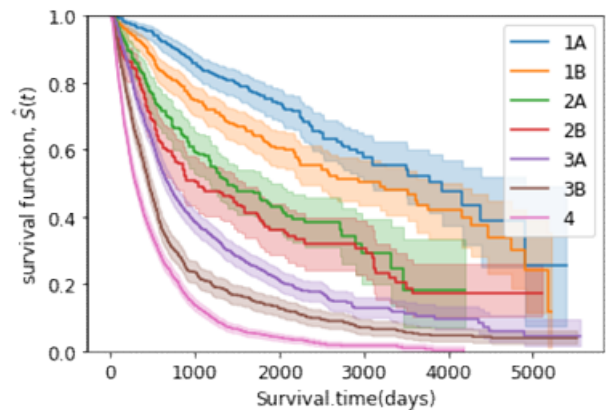


그림 3. 최종병기 카플란마이어 생존 곡선

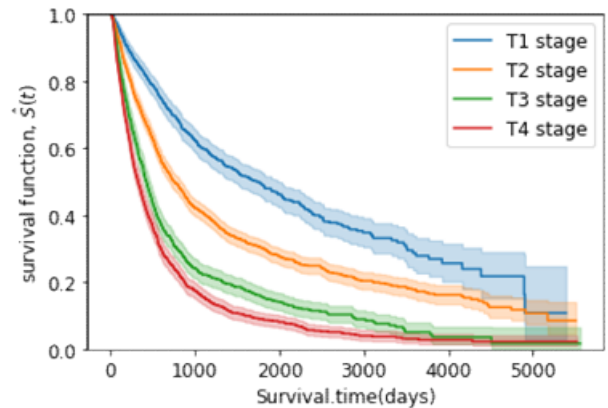


그림 4. T Stage 카플란마이어 생존 곡선

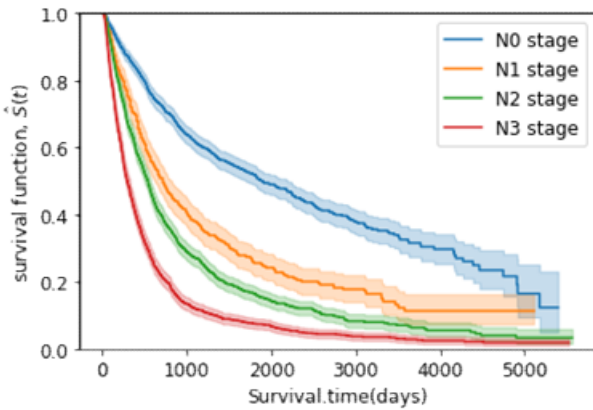


그림 5. N Stage 카플란마이어 생존 곡선

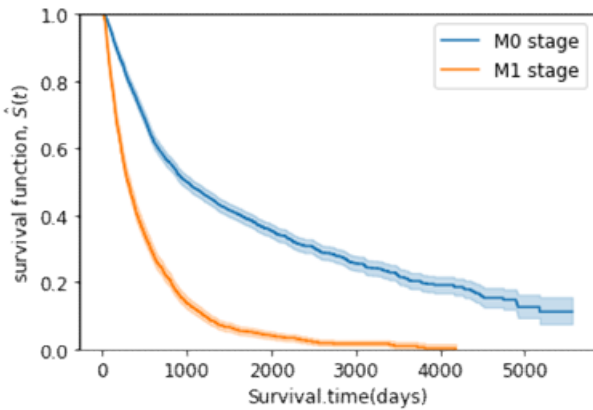


그림 6. M Stage 카플란마이어 생존 곡선

3. Cox비례위험모형 생존분석 성능 비교분석

폐암환자의 TNM병기, 최종병기, 클러스터링 결과를 바탕으로 Cox비례위험모형을 사용하여 폐암환자의 기수별 생존율을 예측하고 c-index를 통해 생존시간 예측정확도를 비교분석한다. 표6과 표7은 생존시간 예측정확도를 비교분석한 결과다. 표6과 표7의 첫 번째 열의 괄호는 클러스터의 수를 의미한다.

표 6의 최종병기4와 최종병기7은 0.680과 0.689로 비슷한 성능을 보인다. 반면에 K를 4와 7로 설정하고 진행한 클러스터링의 경우 모든 결과가 0.5~0.6 사이에 근접하며 무작위로 생존시간을 예측하였다. T, N, M 병기만을 사용했을 때 c-index값이 0.72로 가장 높은 예측 정확도를 보였다. AJCC의 TNM병기 분류표에 의해 결정된 최종병기7보다도 높은 정확도를 보이며 AJCC의 병기분류표가 폐암환자의 생존일을 예측하는데 한계가 있음을 알 수 있다.

AJCC의 7가지 최종병기 분류만으로는 생존분석을 진행하는데 한계가 있기 때문에 K-Means, Agglomerative, Mean-Shift 클러스터링 알고리즘에 K값을 정하지 않고 최적의 생존시간 예측정확도를 구하는 K값을 찾도록 하였다. 표 7

표 6. c-index 생존시간 예측정확도 성능 비교

분류	생존시간 예측정확도 (c-index)
TNM병기	0.704
최종병기4	0.680
최종병기7	0.689
K-Means(4)	0.580
K-Means(7)	0.588
Agglomerative(4)	0.495
Agglomerative(7)	0.499

표 7. c-index 생존시간 예측정확도 성능 비교

분류	생존시간 예측정확도 (c-index)
K-Means(12)	0.642
Agglomerative(3)	0.675
Mean-Shift(4)	0.665

은 알고리즘별 가장 높은 생존시간 예측정확도를 비교분석한 결과다.

클러스터 수가 K-Means는 12개, Agglomerative는 3개, Mean-Shift는 4개일 때 가장 높은 정확도를 보였고 AJCC의 최종병기7 0.689와 비교했을 때 Agglomerative(K=3) 0.675로 약 0.01에 가까운 근소한 차이를 보였다. 각 클러스터링의 군집 결과를 확인해보면 AJCC는 M병기가 1이면 폐암 4기 환자로 결정하는 것과 달리 M병기가 1이어도 2가지 또는 3가지로 나누어 군집되며 c-index 성능도 AJCC의 최종병기와 비슷하다.

종합적으로 보았을 때 현재 AJCC의 TNM병기 분류표에 의한 7가지 최종병기는 폐암환자의 생존일을 예측하는데 적합하지 않으며 T병기, N병기, M병기를 통해 폐암환자의 생존일 예측 할 수 있는 새로운 병기 분류표가 필요하다.

VI. 결론

병원에서는 환자의 3가지 특성 T병기, N병기, M병기를 바탕으로 AJCC에서 제공하는 TNM병기 분류표에 의해 폐암환자의 최종병기를 결정하고 있다. AJCC의 최종병기는 환자의 치료방침, 예후분석, 생존일 예측 등 다양한 분야에서 사용하기에 한계가 있다. 이 논문에서는 일반적인 통계방법을 사용하여 TNM병기 분류의 타당성을 검증하는 연구들과 달리 머신러닝을 활용하여 데이터 과학적 접근을 통해 TNM병기 분류를 평가하였다. 폐암환자의 TNM병기가 생존일에 미치는 영향을 평가하기 위해 TNM병기를 클러스터링하고 AJCC의

최종병기와 함께 Cox비례위험모형과 카플란마이어 생존곡선을 사용하여 비교분석하였다. TNM병기 클러스터링 기반의 생존분석의 성능은 Agglomerative(K=3) 알고리즘을 3개의 스테이지로 설정했을 때 0.675, K-Means (K=12) 알고리즘을 12개의 스테이지로 설정했을 때 0.642로 최종병기 7단계를 기반으로 하는 생존분석과 비슷한 성능을 보였다. 특히, 폐암 환자의 M Stage가 1이면 4기로 분류하는 AJCC의 병기분류와는 다르게 K-Means와 Agglomerative 알고리즘은 4기 환자군을 2~3개의 군으로 분류하고 있다. 즉, M stage를 0과 1로 구분하지 않고 세분화하면 환자의 생존일 예측 성능이 향상될 수 있을 것으로 분석된다.

향후, 폐암 환자의 T, N, M 병기 정보뿐만 아니라 CT 및 PET 이미지를 함께 고려하여 환자의 생존일을 예측 할 수 있는 새로운 병기 분류 기법을 연구하고자 한다.

REFERENCES

- [1] 박예민, 김명건, et al, “가천의대 길병원에서 진단된 폐암 환자의 임상적 특성,” *대한내과학회지*, 제 78권, 제2호, 215-221쪽, 2010년
- [2] Wu, Dengyuan, et al, “An examination of TNM staging of melanoma by a machine learning algorithm,” *2012 International Conference on Computerized Healthcare (ICCH). IEEE*, pp. 120-126, Dec. 2012.
- [3] 김혜영, “폐암의 병기 결정,” *대한의사협회 대한의사협회지*, 제51권, 제12호, 1118-1124쪽, 2008년
- [4] McGhan, Lee J., et al, “Validation of the updated 7th edition AJCC TNM staging criteria for gastric adenocarcinoma,” *Journal of Gastrointestinal Surgery*, vol. 16, no. 1, pp. 53-61, 2012.
- [5] Detterbeck, Frank C., et al, “The IASLC Lung Cancer Staging Project: methodology and validation used in the development of proposals for revision of the stage classification of NSCLC in the forthcoming (eighth) edition of the TNM classification of lung cancer,” *Journal of thoracic oncology*, vol. 11, no. 9, pp. 1433-1446, 2016.
- [6] Chansky, Kari, et al, “The IASLC lung cancer staging project: external validation of the revision of the TNM stage groupings in the eighth edition of the TNM classification of lung cancer,” *Journal of Thoracic Oncology*, vol. 12, no. 7, pp. 1109-1121, 2017.
- [7] Sui, Xizhao, et al, “Validation of the stage groupings in the eighth edition of the TNM classification for lung cancer,” *Journal of Thoracic Oncology*, vol. 12, no. 11, pp. 1679-1686, 2017.
- [8] Goldstraw, Peter, et al, “The IASLC Lung Cancer Staging Project: proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edition of the TNM Classification of malignant tumours,” *Journal of thoracic oncology*, vol. 2, no. 8, pp. 706-714, 2007.
- [9] Goldstraw, Peter, et al, “The IASLC lung cancer staging project: proposals for revision of the TNM stage groupings in the forthcoming (eighth) edition of the TNM classification for lung cancer,” *Journal of Thoracic Oncology*, vol. 11, no. 1, pp. 39-51, 2016.
- [10] 최철웅, 벨무루간 아레수 발라지, 김경백, “TNM 병기 클러스터링을 활용한 폐암 최종병기 결정,” *한국스마트미디어학회 춘계학술대회*, 554-557쪽, 조선대학교, 대한민국, 2020년 5월
- [11] Kaur, Harjot, and Prince Verma, “Comparative Weka Analysis of Clustering Algorithm's,” *International Journal of Information Technology and Computer Science*, pp. 56-67, Aug. 2017.
- [12] 홍의석, 박미경, “대표적인 클러스터링 알고리즘을 사용한 비감독형 결합 예측 모델,” *정보처리학회 논문지/소프트웨어 및 데이터 공학*, 제3권, 제2호, 57-64쪽, 2014년
- [13] Lee, Seungmin, and Beom Park, “Analyzing Brain Cancer Survivability Using SEER (Surveillance, Epidemiology, and End Results Program) Data with method of Kaplan Meier,” *대한인간공학회 학술대회논문집*, 92-96쪽, 2015년
- [14] In, Junyong, and Dong Kyu Lee, “Survival analysis: Part I—analysis of time-to-event,” *Korean journal of anesthesiology*, 제71권, 제3호, 182-191쪽, 2018년
- [15] Steck, Harald, et al, “On ranking in survival analysis: Bounds on the concordance index,” *Advances in neural information processing systems*, 2008.
- [16] Amigó, Enrique, et al, “A comparison of extrinsic clustering evaluation metrics based on formal constraints,” *Information retrieval*, vol. 12, no. 4, pp. 461-486, 2009.
- [17] 이동수, 김영광, 신혜주, 김진술, “K-평균 군집화 알고리즘 및 딥러닝 기반 군중 집계를 이용한 전염병 확진자 접촉 가능성 여부 판단 모니터링 시스템 제안,” *스마트미디어저널*, 제9권, 제3호, 122-129쪽, 2020년 09월
- [18] 임명진, 김판구, 신주현, “리뷰의 의미적 토픽 분

류를 적용한 감성 분석 모델,” *스마트미디어저널*, 제9권, 제2호, 69-77쪽, 2020년 06월

- [19] 이진우, 모하메드 바니아타, 홍지만, “무선 센서 네트워크를 위한 효율적인 클러스터링 기법,” *스마트미디어저널*, 제6권, 제4호, 24-31쪽, 2017년 12월

저 자 소 개



최철웅(정회원)

2014년 조선대학교 컴퓨터통계 학과
학사 졸업

2016년 조선대학교 소프트웨어융합공
학과 석사 졸업

2020년~현재 전남대학교 인공지능융
합학과 박사과정

<주관심분야 : 인공지능, 자연어처리,
분산시스템, 빅데이터 플랫폼, 의료 빅데이터 분석>



김경백(정회원)

1999년 한국과학기술원(KAIST) 학사
졸업

2001년 한국과학기술원(KAIST) 석사
졸업

2007년 한국과학기술원(KAIST) 박사
졸업

2007년~2011년 University of California
Irvine, 박사후연구원

2012년~현재 전남대학교 인공지능융합학과 교수
<주관심분야 : 소프트웨어 정의 네트워크/인프라, 빅데
이터 플랫폼, 그리드/클라우드 네트워크 시스템, 소셜
네트워킹 시스템, 인공지능 적용 가상물리시스템, 블록
체인, 의료 빅데이터 분석>