https://doi.org/10.21186/IPR.2022.7.1.001

생존분석에서의 기계학습

백재욱
한국방송통신대학교 통계 · 데이터과학과

Machine learning in survival analysis

Jaiwook Baik
Department of Statistics · Data Science, Korea National Open University

요 약 본 논문은 중도중단 데이터가 포함된 생존데이터의 경우 적용할 수 있는 기계학습 방법에 대해 살펴보았다. 우선 탐색적인 자료분석으로 각 특성에 대한 분포, 여러 특성들 간의 관계 및 중요도 순위를 파악할 수 있었다. 다음으로 독립변수에 해당하는 여러 특성들과 종속변수에 해당하는 특성(사망여부) 간의 관계를 분류문제로 보고 logistic regression, K nearest neighbor 등의 기계학습 방법들을 적용해본 결과 적은 수의 데이터이지만 통상적인 기계학습 결과에서와 같이 logistic regression보다는 random forest가 성능이 더 좋게 나왔다. 하지만 근래에 성능이 좋다고 하는 artificial neural network나 gradient boost와 같은 기계학습 방법은 성능이 월등히 좋게 나오지 않았는데, 그 이유는 주어진 데이터가 빅데이터가 아니기 때문인 것으로 판명된다. 마지막으로 Kaplan-Meier나 Cox의 비례위험모델과 같은 통상적인 생존분석 방법을 적용하여 어떤 독립변수가 종속변수 (t_i, δ_i)에 결정적인 영향을 미치는지 살펴볼 수 있었으며, 기계학습 방법에 속하는 random forest를 중도중단 데이터가 포함된 생존데이터에도 적용하여 성능을 평가할 수 있었다.

주제어 생존데이터, 기계학습, 분류, 생존분석, 랜덤포레스트

Abstract We investigated various types of machine learning methods that can be applied to censored data. Exploratory data analysis reveals the distribution of each feature, relationships among features. Next, classification problem has been set up where the dependent variable is death_event while the rest of the features are independent variables. After applying various machine learning methods to the data, it has been found that just like many other reports from the artificial intelligence arena random forest performs better than logistic regression. But recently well performed artificial neural network and gradient boost do not perform as expected due to the lack of data. Finally Kaplan-Meier and Cox proportional hazard model have been employed to explore the relationship of the dependent variable (t_i, δ_i) with the independent variables. Also random forest which is used in machine learning has been applied to the survival analysis with censored data.

Key Words survival data, machine learning, classification, survival analysis, random forest

* 이 논문은 2020년 한국방송통신대학교 학술연구비지원을 받아 작성된 것임
Received 04 Jan 2022, Revised 11 Jan 2022
Accepted 20 Jan 2022
Corresponding Author: Jaiwook Baik
(Korea National Open University)
Email: jbaik@knou.ac.kr
ISSN: 2466-1139(Print)
ISSN: 2714-013X(Online)

1. 서론

다양한 데이터 수집 및 빅데이터 기술의 발달로 인해 사회의 여러 분야에서 각종 데이터를 수집하고 장기간에 걸친 관측치를 활용하여 시스템을 모니터링할 수 있게 되었다. 대부분 실제 활용되는 예에서 모니터링하는 주요 목적은 관심의 대상인 특정 사건의 발생여부 또는 발생시간을 더욱 잘 추정하고자 하는 것이다. 그런데 이런 사건발생 데이터에서 주요 과제 중 하나는 어떤 인스턴스는 관측기간 동안 관심사건이 발생하지 않아 중도중단(censored) 데이터가 존재한다는 것이다. 예를 들어 연구기간이 제한되었기 때문에 또는 관측기간 중 개체가 추적되지 않아 일부 인스턴스는 정확한 사건발생시간이 관측되지 않는다. 이와 같이 중도중단 데이터가 존재하는데도 불구하고 이를 무시하고 전통적인 통계적인 방법이나 기계학습 방법을 사용하여 추정을 하는 것은 적절하지 않다.

한편, Kaplan-Meier나 비례위험모델과 같은 생존분석 방법은 생존분석에 적용되는 통계적 방법론의 중요한 한 분야로서 여러 분야에서 나오는 중도중단 데이터를 모델링할 때 발생하는 문제를 적절히 처리할 수 있다. 생존분석 방법을 섭렵한 초기 논문 중 하나로 Chung 등[1]을 들 수 있는데, 이들은 범죄학에서 교도소에 갇혔던 사람이 풀려나서 재범까지 걸리는 시간을 예측하였다. 생존분석 방법에 관해 현존하는 대부분의 책들, 예를 들어 Kelinbaum and Klein[2]은 기계학습 관점보다는 전통적인 생존분석의 관점에서 데이터를 해석하고 있다.

근래에는 중도중단 데이터가 포함된 생존데이터에도 기계학습을 적용할 수 있는 기법들이 개발되었다[3]. Cruz and Wishart[4]와 Kourou 등[5]은 암의 예측에 여러 가지 기계학습 방법들을 비교하고 있다. 본 연구에서는 생존분석을 위해 전통적인 통계적 방법과 함께 다양한 기계학습 방법들에 대해 포괄적으로 살펴본다. 또한 생존분석 연구에서 통상 사용되는 평가지표와 기계학습 분야에서 사용되는 평가지표에 대해서도 알아본다.

본 논문은 다음과 같은 내용으로 구성되어 있다. 2절에서는 본 논문에서 살펴볼 데이터를 소개하고 이에 대한 탐색적 자료분석을 실시해본다. 3절에서는 중도중단 데이터가 포함된 생존데이터인데도 불구하고 이를 무시하고 분류문제로 보고 실시하는 기계학습 방법에 대해 알아본다. 4절에서는 중도중단 데이터를 고려하는 통상적인 생존분석 방법은 물론 기계학습 방법을 적용해보고, 5절에서 논의를 정리하면서 추후과제를 살펴본다.

2. 생존데이터와 탐색적 자료분석

심혈관계 질환은 전 세계적으로 사망원인 1위로 매년 1790만 명이 사망하며, 전체 사망의 31%를 차지한다[6]. 심부전은 심혈관계 질환으로 인해 주로 야기되는 사건으로서 본 논문에서 살펴볼 데이터는 Kaggle에서 주최한 경진대회에서 제시한 데이터로 심부전에 의한 사망을 예측하는 데 사용될 수 있는 299명에 대한 13개의 특성(feature)을 포함하고 있다[7]. 경진대회에서는 심부전으로 인해 야기되는 사망을 예측할 수 있는 모델을 수립하는 것이 목적이었다. 이에 사망 여부의 특성을 나타내는 결과변수는 death_event로 잡았으며, 이의 설명변수라고 할 수 있는 특성은 age(나이), anaemia(빈혈), creatinin_phosphokinase(혈액 내 CPK 효소 수치), diabetes(당뇨병), ejection_fraction(심박출률), high_blood_pressure(고혈압), platelets(혈소판 수치), serum_creatinine(혈청 크레아티닌 수치), serum_sodium(혈청 나트륨 수치), sex(성별), smoking(흡연여부), time(앞의 결과변수인 사망 여부의 판단시간으로 추적기간)으로 총 12개이다.

본 데이터는 전통적인 생존분석 방법인 Kaplan-Meier의 비모수적 방법이나 Cox 비례위험의 반모수적 방법으로 time에 따른 생존확률을 구할 수 있다. 아울러 똑같은 데이터에 대해 12개의 설명변수의 변화에 따라 결과가 death_event의 두 가지로 분류된다고 보고 logic regression, decision tree, random forest, gradient boosting 등과 같은 통상적인 기계학습 방법을 적용하여 적절한 모델은 무엇인지 탐색할 수 있다[8].

일반적으로 여러 가지 특성들이 데이터의 형태로 주어지는 경우 이들 간의 상관관계가 얼마나 큰지 먼저 살펴본다. Figure 1은 13개의 특성에서 두 개씩 짝을 지어 두 특성 간 상관관계가 얼마나 큰지 나타내는 heatmap이다. 두 특성이 age와 platelets와 같이 둘 다 연속형 데이터라면 Pearson 상관계수를 구하고, 두 특성이 모두 sex와 diabetes와 같이 범주형 데이터라면 카이제곱 통계를 사용하는 Phi coefficient를 구하며, 두 특성 중 하나는 death_event와 같이 범주형 데이터이고 다른 하나는 age와 같이 연속형 데이터라면 Pearson 상관계수의 특별한 경우라고 할 수 있는 point biserial correlation coefficient를 구한다.

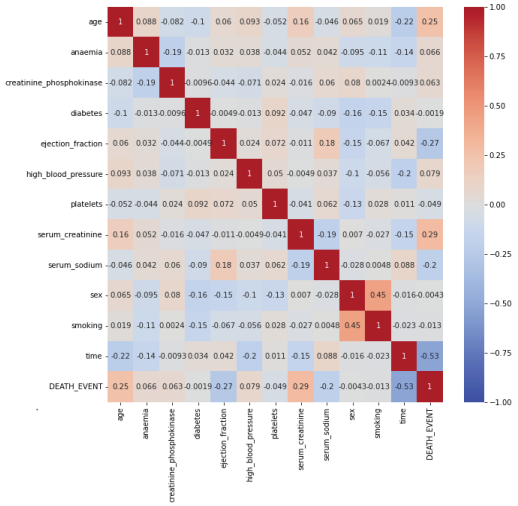

Figure 1. 두 특성 간 상관관계를 나타내는 heatmap

Figure 1의 heatmap에서 색깔이 진한 것은 상관관계가 높은 것을 나타낸다. Python을 이용하여 heatmap을 구하면 파란색과 빨간색은 각각 음과 양의 상관관계를 나타낸다. 따라서 Figure 1의 대각선에서와 같이 똑같은 특성 간의 상관관계는 1이므로 heatmap에서도 가장 진한 빨간색으로 표시된다.

결과변수인 특성 death_event와 밀접한 관련이 있는(임으로 상관계수가 0.2 이상인 것으로 잡았음) 원인변수인 특성은 크기 순서대로 time, serum_creatinine, ejection_fraction, age, serum-sodium이며, 이때 상관계수는 각각 -0.53, 0.29, -0.27, 0.25, -0.2이다.

앞에서와 같은 수치 요약이외에 그림을 이용해도 여러 가지 정보를 얻을 수 있다. 예를 들어, 각 특성의 분포를 그림으로 그려 특이점은 없는지 살펴볼 수 있다. 또한 어떤 특성과 다른 특성(들) 간의 관계를 그림으로 그려 서로 어떤 관련이 있는지 파악할 수도 있다. 본 데이터의 경우에는 개체들의 나이(age)의 분포, 성(sex)별 나이의 분포, 성별 그리고 사망(death_event)여부별 나이의 분포에 대해 관심이 있을 수 있다. Figure 2, Figure 3 및 Figure 4는 각각의 상황을 나타낸다. Figure 2로부터 나이가 40세 미만 95세 이상인 사람은 없고, 60세가 제일 많으며, 45세부터 70세 사이에 많이 분포한다는 것을 알 수 있다. Figure 3으로부터는 남녀 모두 중위수가 60세이지만 남자가 여자보다 고령자가 더 많다는 것을 알 수 있다. Figure 4는 남녀별로 그리고 사망여부별로 나이의 분포를 나타낸다. 여자와 달리 남자의 경우 고령자는 생존자 쪽보다는 사망자 쪽에서 더 많음을 알 수 있다.

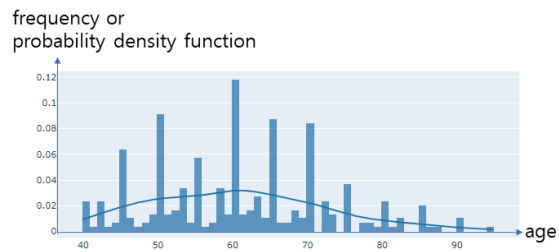

Figure 2. Frequency or probability density function of age

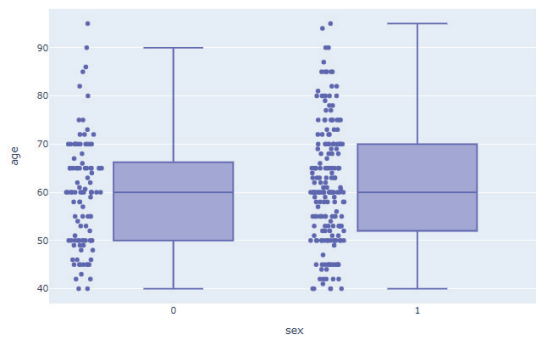

Figure 3. Distribution of age by sex(sex 0: 여자, sex 1: 남자)

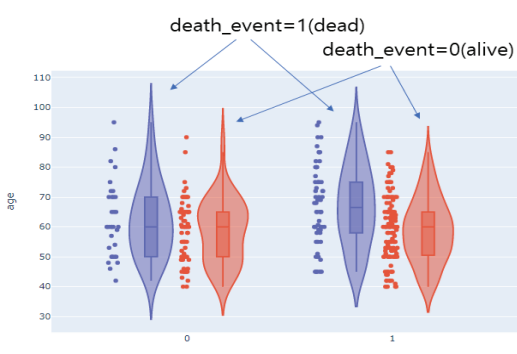

Figure 4. Distribution of age by sex and death_event

3. 분류문제의 기계학습 방법

이 절에서는 age, anaemia 등의 원인변수인 12개의

특성들과 결과변수인 특성 death_event와의 관계를 기계학습 방법을 통해 살펴보고자 한다. 여기서 death_event는 1(사망)과 0(생존)의 두 가지 범주를 가지므로 중도중단 데이터가 포함된 데이터인데도 불구하고 단지 분류문제로 보고 기계학습을 한 것이다.

기계학습은 병원기록 데이터에 적용하여 심부전 증상을 가지고 있는 환자의 생존을 예측하고[9], 심부전을 불러일으키는 임상적 특성이 무엇인지 밝히는[10] 매우 효율적인 도구이다. 따라서 데이터과학자는 임상에서의 예측뿐만 아니라[11] 여러 특성들의 중요 순위의 결정에도 기계학습을 활용한다. 앞의 기계학습보다 좀 더 나은 예측력을 가진 인공지능이 되려면 의료기록정보의 활용[12]은 물론 영상정보[13]의 활용도 필요하다. 근래에는 심층학습과 메타분석이 의학 분야에 적용되고 있으며[14], 아직 정확도는 낮은 상태이지만 인간 전문가의 직무를 향상시켜주고 있다[15].

Chicco and Juman[8]은 2절의 데이터에 대해 여러 가지 기계학습 방법을 적용하여 환자의 생존여부를 예측하였고 여러 가지 특성들을 중요도 면에서 순서를 매겼다. 그 결과 ejection_fraction과 serum_creatinine의 두 특성만 모델에 적용해도 높은 예측력을 달성할 수 있었다. 따라서 본 연구에서는 독립변수로 ejection_fraction, serum_creatinine 및 time을 잡고, 이들이 종속변수인 death_event를 예측하는 여러 기계학습 방법에 대해 살펴보기로 한다.

기계학습 방법에서는 전체 데이터를 모델 수립에 사용되는 training 데이터와 모델의 성능을 평가하는 test 데이터로 나눈다. 본 연구에서는 전체 299개의 데이터 중 80%에 해당하는 데이터는 training 데이터로, 나머지 20%에 해당하는 데이터는 test 데이터로 나눈다. 또한 독립변수들에 대해 sklearn의 StandardScaler를 활용하여 평균이 0이고 표준편차가 1인 값들로 변환한다.

기계학습 방법 중 logistic regression은 종속변수가 두 개의 범주로 나누어지는 경우 독립변수들의 영향력을 파악하기 위해 적용하는 통계적 분석방법이다. 주어진 데이터의 경우 training 데이터를 활용하여 모델을 세우고 test 데이터에 대해 모델의 성능을 평가한 결과 Table 1과 같은 confusion matrix를 구할 수 있었다. 따라서 실제로 사망한 37명의 사람들 중 사망할 것이라고 제대로 예측한 true positive(TP)는 34명이고 잘못 예측한 false negative(FN) 3명이다. 한편, 실제로 생존한 23명의 사람들 중 생존할 것이라고 제대로 예측한 true negative(TN)는 12명에 불과하고 잘못 예측한 false positive(FP)는 11명이다. 따라서 정확도(accuracy)는 (34+12)/(34+3+11+12)=0.77로 낮은 편이다.

Table 1. confusion matrices for various types of machine learning

			predicted values	
			positive (1)	negative (0)
actual values	logistic regression	positive (1)	34	3
		negative (0)	11	12
	K nearest neighbor	positive (1)	35	2
		negative (0)	5	18
	support vector machine	positive (1)	35	2
		negative (0)	8	15
	decision tree	positive (1)	33	4
		negative (0)	4	19
	random forest	positive (1)	36	1
		negative (0)	6	17
	artificial neural network	positive (1)	34	3
		negative (0)	8	15
	XGBoost	positive (1)	36	1
		negative (0)	7	16

앞의 logistic regression에서와 같이 K nearest neighbor, support vector machine, decision tree, random forest, artificial neural network 및 Gradient boost의 방법을 이용하여 주어진 데이터에 대해 모델을 설정할 수 있고, 그 성능을 정확도 측면에서 구할 수 있다. Table 1은 그 결과를 보여준다. 지금까지의 기계학습 방법들을 주어진 데이터에 적용한 결과 성능을 비교하면 Figure 5에서와 같다.

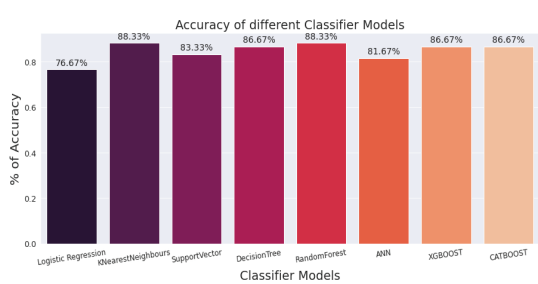

Figure 5. 여러 기계학습 모델의 성능: 정확도

4. 중도중단 데이터를 고려한 생존분석 방법

앞의 3절에서는 종속변수로 사망여부(death_event)만을 고려했다. 하지만 사망여부는 추적시간(time)과 밀접한 관련이 있다. 따라서 이 두 개의 특성을 함께 고려하는 변수를 종속변수로 두고 분석할 필요가 있다. 전통적인 생존분석에서는 i번째 개체의 시간을 t_i, 중도중단 여부를 δ_i로 나타내어, i번째 개체가 시간 t_i에 사망했으면 δ_i=1로 두고, 시간 t_i까지는 살아있었고 더 이상 추적이 되지 않아 중도중단된 상태이면 δ_i=0으로 둔다.

4.1 Kaplan-Meier 방법

Kaplan-Meier 방법은 추적시간과 중도중단 여부 (t_i, δ_i)만을 고려하여 비모수적인 방법으로 시간이 흐름에 따라 생존확률이 어떻게 떨어지는지 보여준다. 특히 어떤 특정 독립변수의 값을 기준으로 데이터를 그룹화하지 않으면 전체 데이터의 평균적인 경향만을 보여준다. Figure 6에는 주어진 데이터의 경우 시간이 흐름에 따라 Kaplan-Mier 방법에 의한 생존확률이 어떻게 감소하는 지 보여준다.

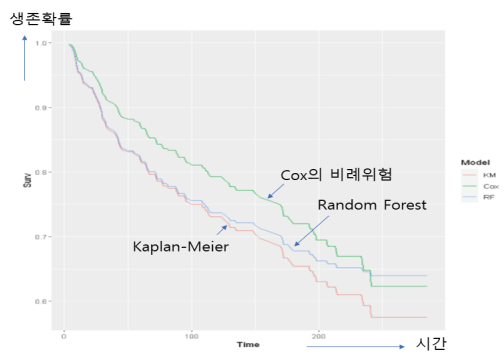

Figure 6. 시간에 따른 생존확률의 추이(3가지 방법)

하지만 특정 독립변수가 생존확률에 미치는 영향력을 보기 위하여 해당 독립변수의 특정한 값을 기준으로 데이터를 그룹핑할 수 있다. 예를 들어 심박출률(ejection_fraction)이 높은 사람과 낮은 사람 간에 생존가능성에 차이가 있는지 보기 위하여, 예를 들어 평균 심박출률인 38.1%를 기준으로 이 보다 높은 사람과 낮은 사람의 두 그룹으로 나눌 수 있으며, 각 그룹에 대한 시간에 따른 생존확률의 추이를 Kaplan-Meier 방법으로 살펴볼 수도 있다.

4.2 Cox proportional hazards 모델

Cox proportional hazards(Cox의 비례위험) 모델은 생존분석 분야에서 독립변수들이 종속변수에 미치는 영향력을 평가하는데 많이 사용되는 기법이다. Cox의 비례위험 모델은 다음과 같이 쓸 수 있다.

$$h(t,x) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)$$

여기에서 $h(t,x)$는 독립변수 $x_1, x_2, ,..., x_p$를 고려했을 때의 위험함수(hazard function)이며 $h_0(t)$는 $x_1 = x_2 = ... = x_p = 0$일 때의 기저위험함수(baseline hazard function)이다. 비례위험 모델의 특징은 기저위험함수는 다른 공변수에 의존하지 않고 단지 시간 t에만 영향을 받는다는 것이다. 또한 공변수는 위험함수에 비례적으로 작용한다는 것이다.

종속변수 (t_i, δ_i)에 영향을 줄 수 있는 11개 독립변수들의 영향력을 살펴보기 위해 주어진 데이터에 대해 Cox의 비례위험 모델을 적합시킨 결과 Table 2를 얻었다. 이로부터 종속변수 (t_i, δ_i)에 영향을 크게 미치는 독립변수는 age, ejection_fraction, serum_creatinine 등이라는 것을 알 수 있다.

Table 2. Cox의 비례위험 모델 적용 결과

Call:
coxph(formula = Surv(time, death_event) ~ age + anaemia + cre_pho + diabetes + eje_fra + hi_blo + platelets + ser_cre + ser_sod + sex + smoking)
n= 299, number of events= 96

	coef	exp (coef)	se (coef)	z	Pr(>|z|)
age	4.61e-02	1.05e+00	9.32e-03	4.977	6.45e-07 ***
anaemiayes	4.60e-01	1.58e+00	2.16e-01	2.122	0.0338 *
cre_pho	2.20e-04	1.00e+00	9.91e-05	2.225	0.0260 *
diabeteyes	1.39e-01	1.15e+00	2.23e-01	0.627	0.5307
eje_fra	-4.89e-02	9.52e-01	1.08e-02	-4.672	2.98e-06 ***
hi_bloyes	4.75e-01	1.61e+00	2.16e-01	2.201	0.0278 *
platelets	-4.65e-07	1.00e+00	1.12e-06	-0.412	0.6806
ser_cre	3.21e-01	1.38e+00	7.01e-02	4.575	4.76e-

					06 ***
ser_sod	-4.419e-02	9.568e-01	2.327e-02	-1.899	0.0575 .
sexmale	-2.375e-01	7.886e-01	2.516e-01	-0.944	0.3452
smokingyes	1.289e-01	1.138e+00	2.512e-01	0.513	0.6078

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

앞의 Cox의 비례위험 모델에서는 11개의 독립변수가 종속변수 (t_i, δ_i)에 미치는 영향력을 모두 고려했으며, 이 경우 시간에 따라 생존확률이 어떻게 변하는지 관심이 있을 수 있다. Figure 6에는 Cox의 비례위험모형에서 R의 survfit을 이용하여 시간의 흐름에 따른 생존확률의 추이를 살펴볼 수 있다. 주어진 데이터의 경우 Cox의 비례위험 모델로부터 나온 생존확률 추정치가 Kaplan-Meier 방법에 의한 생존확률 추정치보다 더 큰 것으로 보인다. 예를 들어 100시간에서의 생존확률은 Kaplan-Meier 방법으로는 75%이지만 Cox의 비례위험 모델로는 81% 정도이다.

4.3 random survival forest 모델

random survival forest(RSF)는 생존데이터를 분석하기 위해 random forest를 발전시킨 것으로 random forest의 알고리즘을 기본으로 하며**[16]**, 다음 절차를 밟는다.

1단계: 주어진 자료로부터 B(b=1, 2, ..., B)개의 bootstrap sample을 생성한다. 생성한 boostrap sample 중에서 일부를 out-of-bag(OOB)라 칭하고, 이들을 제외한 sample(in-bag boostrap sample)로 모형을 생성한다.

2단계: in-bag bootstrap sample로 survival tree를 성장시킨다. 이때 각 마디에서 p개의 후보변수를 랜덤하게 골라, 이 중에서 자식 마디의 동질성이 최대가 되는 변수를 선택하고, 최적 분리가 발생하는 지점을 찾는다. 분리규칙은 log-rank 검정통계량을 많이 이용하며, stopping criterion에 도달할 때까지 마디를 분리해나간다.

3단계: 마디가 더 이상 분리되지 않는 지점에 도달하면 그 마디를 끝마디(terminal node)라 하고, tree의 끝마디들에서 얻은 정보를 결합하여 앙상블 예측모형을 얻는다.

R에서는 ranger라는 함수를 이용하여 중도중단 데이터가 포함된 생존데이터에 대해 RSF 모델을 적합시킬 수 있다. ranger에서는 따로 지정하지 않으면 500개의 survival tree를 생성하며, 각 마디에서는 남은 변수들의 수에 루트를 취한 수만큼의 변수들 중에서 분기가 가장 잘 되는 변수를 선택한다. 주어진 데이터의 경우 독립변수 11개 모두를 고려하여 RSF 모델을 적용한 결과 Table 3과 같은 결과를 얻었다. Figure 6에는 RSF 모델을 적용한 결과 시간에 따른 생존확률의 추이를 보여준다.

Table 3. RSF 모델 적용 결과

Ranger result
Call:
ranger(Surv(time, death_event) ~ age + anaemia + cre_pho + diabetes + eje_fra + hi_blo + platelets + ser_cre + ser_sod + sex + smoking, data = heart, mtry = 4, importance = "permutation", splitrule = "extratrees", verbose = TRUE)

Type:	Survival
Number of trees:	500
Sample size:	299
Number of independent variables:	11
Mtry:	4
Target node size:	3
Variable importance mode:	permutation
Splitrule:	extratrees
Number of unique death times:	148
Number of random splits:	1
OOB prediction error (1-C):	0.2939115

5. 결론

본 논문은 중도중단 데이터가 포함된 생존데이터의 경우 기계학습 방법에 대해 살펴보았다. 우선 탐색적 자료분석으로 주어진 데이터에 대한 특별한 가정이 없이 그림으로 각 특성에 대한 분포를 파악할 수 있었고, 여러 특성들 간의 관계 또한 파악할 수 있었으며, 종속변수에 해당하는 특성과 독립변수라고 할 수 있는 특성 간의 상관관계를 통해 종속변수에 영향을 주리라 기대되는 독립변수들의 순위를 결정할 수 있었다.

다음으로 독립변수에 해당하는 특성들이 종속변수에 해당하는 특성에 영향을 미친다고 볼 수 있는데, 종속변수의 특성이 사망 또는 생존으로 나누어지므로 해당 문제를 분류문제로 보고 logistic regression, K nearest

neighbor, support vector machine, decision tree, random forest, artificial neural network, gradient boost의 기계학습 방법을 적용해보았다. 그 결과 주어진 299명의 개체에 대한 데이터의 분석으로는 K nearest neighbor와 random forest가 정확도가 가장 높고, 그 다음으로 decision tree, gradient boost가 높으며, 그 다음으로 support vector machine이 높으며, 그 다음으로 artificial neural network가 높으며, 마지막으로 logistic regression이 가장 낮은 것으로 나왔다. 주어진 데이터의 경우 logistic regression의 정확도가 가장 낮게 나왔고, random forest가 정확도가 가장 높게 나왔는데, 이는 빅데이터의 경우 기계학습 방법을 적용하면 통상적으로 나오는 결과이다. 하지만 인공지능 분야에서 근래 성능이 좋다고 하는 artificial neural network나 gradient boost가 K nearest neighbor나 decision tree보다 훨씬 더 좋은 성능을 보이지 않는데, 그 이유는 주어진 데이터의 수가 299개로 많지 않기 때문인 것으로 판명된다.

앞의 기계학습 방법에서는 개체의 사망여부(death_event)만을 종속변수로 보았는데, 개체의 사망여부와 사망여부의 판단시간인 추적기간(time) 간에는 밀접한 관련이 있다. 따라서 통상적인 생존분석에서와 같이 11개의 독립변수와 개체의 사망여부 및 추적기간을 하나로 묶은 종속변수 (t_i, δ_i) 간의 관계를 살펴볼 필요가 있다. 이에 비모수적 방법인 Kaplan-Meier를 이용하여 개체들의 시간에 따른 생존확률의 추이를 살펴보았으며, 반모수적인 방법인 Cox의 비례위험모형을 이용하여 어떤 독립변수가 중요한지도 살펴보았다. 마지막으로 기계학습 방법에 속하는 random forest를 중도중단 데이터가 포함된 생존데이터에도 적용하여 성능을 평가하였다.

본 연구에서 살펴본 생존데이터는 개체수가 299개에 불과하지만 기계학습 방법을 적용하면 logistic regression보다는 random forest가 더 성능이 뛰어나다는 것을 파악할 수 있었다. 하지만 근래의 기계학습 방법들 중 성능이 뛰어나다고 하는 artificial neural network나 gradient boost와 같은 방법은 성능을 확인할 수 없었다. 그 이유는 본 연구에서 살펴본 데이터의 개체수가 적기 때문인 것으로 판단된다. 따라서 다음 연구에서는 중도중단 데이터가 포함된 빅데이터를 대상으로 여러 가지 기계학습 방법을 적용하여 그 성능을 평가하고자 한다.

생존데이터는 중도중단 데이터를 포함하기도 하지만 때로는 여러 사유로 인해 개체가 사망하는 경우가 많다. 이에 통상적인 생존분석에서는 competing risk model을 적용하는데, 다음 연구에서는 이런 model이 random forest와 같은 기계학습 방법에 어떻게 적용할 수 있는지 살펴보고자 한다. 마지막으로 생존분석에서는 recurrent event가 많이 생길 수 있는데, 기계학습 방법에서는 이런 event를 어떻게 처리하는 것이 합리적인지 알아보고자 한다.

References

[1] Chung, C., Schmidt, P. and Witte, A. (1991). Survival analysis: A survey. Journal of Quantitative Criminology, 7(1), 59-98.

[2] Kleinbaum, D. G. and Klein, M. (2006). Survival analysis: A self-learning text. Springer Science & Business Media.

[3] Wang, P., Li, Y. and Reddy, C. K. (2019). Machine learning for survival analysis: A Survey. ACM computing Surveys, 51(6), 1-36.

[4] Cruz, J. A. and Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. Cancer informatics, 2.

[5] Kourou K. et al. (2015). Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 13, 8 - 17.

[6] WHO. (2016). Fact sheet on CVDs. Gloval Hearts. World Health Organization.

[7] Ahmad, T. et al. (2017). Survival analysis of heart failure patients: A case study. PloS ONE, 12(7).

[8] Chicco, D and Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making 20(16).

[9] Al'Aref S. J. et al. (2019). Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. European Heart

Journal, 40(24), 1975 - 1986.
[10] Dunn W. B. et al. (2007). Serum metabolomics reveals many novel metabolic markers of heart failure, including pseudouridine and 2-oxoglutarate. Metabolomics, 3(4), 413 - 426.
[11] Ambale-Venkatesh B. et al. (2017). Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. Circr Res., 121(9), 1092 - 1101.
[12] Panahiazar M. et al. (2015). Using EHRs and machine learning for heart failure survival analysis. Stud Health Technol Informat., 216, 40-44.
[13] Ahmad T. et al. (2018). Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients. Journal of American Heart Association, 7(8).
[14] Krittanawong C. et al. (2019). Deep learning for cardiovascular medicine: A practical primer. European Heart Journal, 40, 2058 - 2073.
[15] Bello G. A. et al. (2019). Deep learning cardiac motion analysis for human survival prediction. Nature Machine Intelligence, 1, 95-104.
[16] Ishwaran, H. et al. (2008). Random Survival Forests. The Annuals of Applied Statistics, 2, 841-860.

백 재 욱(Baik, Jai Wook)

- 1992년 04월~ 현재 : 한국방송통신대학교 통계· 데이터과학과 교수
- 1986년 09월~ 1991년 05월 : 미국 Virginia Polytechnic Institut and State University 통계학박사
- 1983년 09월~ 1986년 05월 : 미국 University of Wisconsin-Madison 통계학석사
- 1976년 03월~ 1983년 02월 : 중앙대학교 응용통계학과 학사
- 관심분야 : 통계학, 생산관리
- E-Mail : jbaik@knou.ac.kr