Journal of the Korean Data & Information Science Society
2018, **29**(6), 1391–1408

http://dx.doi.org/10.7465/jkdi.2018.29.6.1391
한국데이터정보과학회지

유전체 자료분석을 위한 생존분석방법에 관한 고찰[†]

이승연[1]

[1]세종대학교 수학통계학부

접수 2018년 10월 15일, 수정 2018년 11월 15일, 게재확정 2018년 11월 16일

요 약

관심의 대상이 되는 사건이 발생할 때까지 걸리는 생존시간을 다루는 생존분석의 가장 큰 특성은 생존시간이 완전하게 관측되지 않고 중도절단 된다는 점이다. 이러한 중도절단자료의 특성을 고려하여 추정, 검정 및 모형적합에 대하여 고전적인 생존분석 방법들이 많이 개발되어져 왔으나, 마이크로어레이자료를 시작으로 대용량의 유전체 자료가 수집되면서 유전적 정보와 생존시간과의 연관성 연구가 진행되면서 표본의 수에 비하여 엄청나게 많은 수의 유전정보 변수들을 다루는 새로운 통계적인 방법들이 생존자료에 확장되었다. 결과적으로 기존의 임상자료로만 구축된 통계예측모형에 유전체 정보가 추가적으로 고려됨으로써 생존함수에 대한 예측력이 향상되었고, 개인의 유전정보에 따라 더 적합한 치료방법이나 치료약을 개발해야 한다는 개인맞춤의학의 필요성이 부각되기 시작되었다. 다양한 첨단 생물학 기술을 통하여 서로 다른 형태의 대용량의 유전체 자료를 통합하는 방법론에 대한 연구들이 이루어지면서 기계학습 방법이 생존분석에 접목되어 많은 연구방법들이 개발되고 있다. 본 연구에서는 기존의 임상자료를 기반으로 분석하는 전통적인 생존분석 방법들을 소개하고, 고차원의 유전체 자료를 분석하기 위한 생존분석 방법들과 통합적인 유전체 자료분석을 위하여 생존분석에 접목된 기계학습방법들에 대하여 간략하게 살펴보고자 한다.

주요용어: 기계학습, 벌점함수, 비모수적인 방법, 생존시간, 중도절단, 통계예측모형.

1. 서론

생존분석 (survival analysis)은 어떤 시점에서부터 관심의 대상이 되는 사건이 발생할 때까지 걸리는 생존시간 (survival time)에 대하여 통계적인 추론을 다루는 분야이다. 생존분석에서는 완전하게 관측된 생존시간 외에 중도절단 (censoring)된 자료를 포함하게 되는데 이는 연구 기간 내에 관심의 대상이 되는 사건이 발생하지 않아 실제의 생존시간을 관측하지 못하는 경우가 흔히 일어나기 때문이다. 이와 같이 중도절단자료를 분석해야 하는 것이 생존분석의 특성이며 생존분석의 이론과 방법들은 중도절단자료를 분석할 수 있도록 개발되어 왔다. 임상시험과 같은 코호트 연구의 중요한 목적은 오랜 기간 동안 추적조사로 관측되는 생존시간의 생존함수를 추정하거나 기존의 치료방법과 새로운 치료방법 간의 생존율에 유의한 차이가 나는지 여부를 검정하는 것이다. 또한 생존시간과 연관성이 있는 요인들을 인구통계학적, 임상학적, 환경적인 자료들로부터 찾아내어 통계예측모형을 구축하는 것이다.

전통적인 생존분석에서 다루는 주제는 크게 세 가지로 요약하면 생존함수의 추정, 그룹 간의 생존함수의 비교, 생존시간을 예측할 수 있는 위험요인들과의 연관성 검정 및 통계모형의 구축이라고 할 수 있다.

[†] 이 논문은 2016년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(NRF-2016R1D1A1B03934908)입니다.

[1] (05006) 서울특별시 광진구 능동 209, 세종대학교 수학통계학부, 교수.
E-mail: leesy@sejong.ac.kr

대표적인 생존함수 추정량은 Kaplan-Meier (1958)가 제안한 비모수적인 카플란-마이어 추정량으로 대부분의 생존분석 결과에서 찾아볼 수 있다. 두 그룹 간의 생존곡선을 검정하는 방법으로는 로그-순위 검정법이 가장 많이 활용되는데 이 방법도 비모수적인 검정법이다. 생존시간에 대한 통계예측모형으로는 Cox (1972)가 제안한 칵스모형으로 위험률함수를 기반으로 한 회귀모형이다. 전통적인 생존분석은 나이, 성별, 인종, 사회 계층적 변수와 같은 인구통계학적 요인, 혈압, 콜레스테롤, BMI, 종양크기 등과 같은 임상학적 요인과 흡연여부, 미세먼지 오염도, 오존농도와 같은 환경적 요인들을 기반으로 표본 수에 비하여 변수의 수가 비교적 적은 경우에 대하여 생존율에 대한 예측율을 높이는 통계모형을 개발하는 연구방법들에 주력하였다.

21세기에 들어서서 생물학의 최첨단 기술의 발달로 유전정보에 대한 DNA microarrays 자료(Eisen 등, 1998; Alizadeh 등, 2000; Dudoit 등, 2002; Tusher 등, 2001; Lee 등, 2000; Newton 등, 2000)가 수집된 이후로 SNP (single nucleotide polymorphism), CNV (copy number variant), NGS (next generation sequencing), methylation, protein chips 등 (Sebat, 2007; Korn 등, 2008) 다양한 첨단 생물학 기술을 통하여 유전체 자료가 대용량으로 수집되어 통계적으로 분석될 필요성이 급증하였다. 이러한 유전자 정보를 분석하는 문제는 결국 표본 수에 비하여 엄청나게 많은 변수의 수 (유전자 정보 수)를 갖는 "차원의 저주 (curse of dimensionality)" (Theodoridis와 Konstantinos, 2008)를 풀어야 하는 문제로 직결되었다. 많은 통계학자들이 이 문제의 해결책으로 변수의 회귀계수에 벌점함수(penalty function)를 제한 조건으로 두어 실제로 효과가 큰 유전자들만을 선택할 수 있는 방법들을 개발하였다. 대표적인 벌점함수로는 Lasso, Ridge, Elastic-Net이 있으며 벌점함수를 이용한 방법들이 기존의 칵스모형으로 확장되어 고차원의 유전자 정보를 적합할 수 있는 모형들이 개발되었고, 기존의 임상학적인 요인에 유전정보를 추가함으로써 생존시간에 대한 예측력을 높이는 연구들이 개발되었다.

1990년에 시작하여 2003년도에 종료된 인간게놈 프로젝트 (`www.ornl.gov/sci/techresources/Human_Genome/index.shtml`)에서 한 명의 개인으로부터 30억 개의 DNA 정보를 해독하는데 27억 달러라는 천문학적인 비용이 들었는데 현재는 동일한 작업이 30시간에 1000달러의 비용으로 가능하게 되었다. 개인의 유전정보에 대한 자료는 의사의 진단과 같이 건강상태를 파악할 수 있는 진단마커로 수요가 더욱 증가하는 추세이다. 한 개인에 대한 유전정보도 microarrys, SNP, protein chip, RNA-seq 등 여러 생물학 기술로부터 다른 정보들이 얻어지므로 이와 같이 서로 다른 형태의 유전정보를 효과적으로 통합하여 예측력을 높일 수 있는 통계적 모형을 구축하려는 연구들이 지속적으로 진행되고 있다. 이러한 추세는 동일한 질환이라도 치료약이나 치료방법을 개인의 유전자 정보에 따라 맞춤으로 제공하려는 정밀의학의 개발로 가속화 되고있다. 이와 같이 복잡하고 노이즈가 많은 자료들을 통합하여 일정한 패턴을 찾아내는데 통계적인 확률이론과 모형들을 이용하는 기계학습방법은 축적된 유전체 자료를 분석하고자 하는 생존분석의 수요와 맞물려 이미 20여 년 전부터 암 질환을 진단하는데 활용되기 시작하였다 (Cruz와 Wishart, 2006). 최근에는 암 질환의 위험여부, 재발여부 및 암 질환의 생존율을 예측하기 위하여 artificial neural network (ANN), decision tree (DT), genetic algorithm (GA) 등과 같은 기계학습방법에 기반한 통계적 방법들이 생존분석에서 크게 활용되고 있다. 앞으로는 유전체 정보뿐 만 아니라 전산화된 병원 기록 데이터까지 통합한 자료를 분석하기 위하여 딥러닝 (deep learning)의 기법들이 도입되는 추세로 발전될 것으로 전망된다.

본 연구에서는 생존분석의 기초적인 개념과 이론에 대하여 개괄적으로 소개하고 시대의 흐름에 따라 발전해 온 생존분석 방법들에 대하여 살펴보고자 한다. 먼저 2절에서 생존분석에서 다루는 기초적인 이론과 함수에 대하여 간단하게 소개하고, 3절에서 전통적인 생존분석 방법들을 소개하겠다. 4절에서는 고차원의 유전체 자료를 분석하기 위한 통계적인 방법들에 관하여 살펴보고, 5절에서는 기계학습과 관련된 생존분석 방법들에 대하여 활용되는 예를 중심으로 소개하고 6절에 결론으로 본 논문을 끝맺음 하고자 한다.

2. 생존분석의 기초적 이론과 함수들

2.1. 생존시간과 중도절단

생존시간 (survival time)은 관심의 대상이 되는 사건 (event)이 발생할 때까지 걸리는 시간으로 정의되며 사건은 사망, 질환의 발병 또는 재발, 실업자의 구직, 보험지급 등과 같이 의학, 보건학, 공학과 사회학 전 분야에 걸쳐 관측되는 현상이 포함될 수 있다. 생존시간자료는 현황적 (cross-sectional)으로 한 시점에서 관측되는 자료가 아니라 주로 전향적 (prospective)으로 진행되는 코호트 연구에서 관측된다. 생존시간은 관심의 대상이 되는 사건이 발생할 때까지 걸리는 시간이므로 연구기간이 종료될 때까지 사건이 발생하지 않으면 생존시간을 완전하게 관측할 수 없게 되는데 이러한 현상을 중도절단 (censoring)이라고 한다. 이 경우 중도절단된 자료는 비록 생존시간이 완전하게 관측되지는 않았지만 결측치 (missing data)와는 전혀 다르다. 중도절단된 시점까지 관측된 정보로부터 생존시간이 적어도 중도절단시간 이후에 발생할 것이라는 정보를 주고 있기 때문이다. 이와 같이 생존시간의 하한값을 아는 경우는 우중도절단 (right censoring)이라고 하고 상한값을 아는 경우를 좌중도절단 (left censoring), 하한값과 상한값을 아는 경우를 구간중도절단 (interval censoring)이라고 한다 (Klein과 Moeschberger, 2010). 이러한 중도절단현상은 생존분석에서 중요한 특성으로 중도절단자료의 정보가 통계적인 추론에 반영될 수 있도록 생존분석방법들이 개발되었다.

2.2. 생존함수, 위험률함수와 누적위험률함수

생존분석에서 중요한 목적 중의 하나는 생존함수 (survival function)을 추정하는 것이다. 생존시간을 T 라고 하면 T 는 음이 아닌 연속형 확률변수이며 확률밀도함수를 $f(t)$, 확률누적분포함수를 $F(t)$로 갖는다고 하자. 생존함수는 생존시간 T 가 시간 t 이후에 발생하는 확률로 정의되며 $S(t) = P(T > t) = 1 - F(t)$로 표현된다. 생존분석에서는 확률밀도함수보다도 조건부 확률밀도함수의 개념을 가진 위험률함수 (hazard function)를 기반으로 추론하는 경우가 더 많은데 위험률함수는 생존시간 T 가 시간 t까지 발생하지 않았다는 조건 하에서 구간 $[t, t + \Delta t]$ 에서 순간적으로 발생할 확률로 정의되며 다음과 같이 표현된다.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(T \le T < t + \Delta t | T \ge t)}{\Delta t} = \frac{f(t)}{S(t)}.$$

누적위험률함수 (cumulative hazard function)는 시간 t 까지 누적된 위험률로 정의되면 $H(t) = \int_0^t h(s)ds$로 표현된다. 또한 생존함수, 위험률함수와 누적위험률함수 간에는 서로 다음과 같은 관계식이 성립한다.

$$h(t) = \frac{f(t)}{S(t)} = \frac{-d \log S(t)}{dt}, \quad H(t) = -\log S(t), \quad S(t) = \exp(-H(t)).$$

생존분석에서는 생존시간을 관측하는 과정에서 중도절단되는 자료가 발생하기 때문에 어떤 시점 t까지 사건이 발생하는 확률을 추정하기 어렵다. 대신에 어떤 시점 t 전에 이미 사건이 발생하였거나 중도절단된 자료가 주어진 조건 하에서 생존할 확률을 추정할 수는 있다. 따라서 자연스럽게 어떤 시점 t까지 누적해 온 조건부적인 정보를 기반으로 시점 t 이후에 사건이 발생하는 확률의 개념인 위험률함수를 이용하여 생존함수를 추정하게 된다. 실제로 Kaplan-Meier (1958)의 생존함수 추정량은 관측된 생존시간의 각 시점에서 추정된 조건부 생존확률의 곱으로 구하는데 이 계산과정에서 위험률함수의 추정량을 이용한다. 위의 식에서 나타난 동등한 관계식을 통하여 생존함수, 위험률함수와 누적위험률함수 중 하나를 추정하면 다른 함수도 쉽게 추정할 수 있다. 생존분석에서 다루는 생존시간의 모수적 분포는 양

의 실수 구간에서 분포하며 일반적으로 비대칭적이며 꼬리가 길게 늘어지는 형태를 띠는 경우가 많다. 생존분석에서 많이 활용되는 분포는 지수분포, 와이블 분포, 감마분포, 로그-정규분포, 로그-로지스틱 분포, Gompertz 분포, 파레토 분포 등으로 이 분포의 확률밀도함수, 위험률함수, 생존함수와 기댓값에 대한 정보는 Kim (2016, Table 2.3)를 참조하라.

3. 전통적인 생존분석방법

생존시간과 중도절단시간을 각각 T 와 C 라고 하자. 관측된 시간을 $\tilde{T} = \min(T, C)$라고 하고 $\delta = I(T < C)$라고 할 때 관측된 생존자료는 $(\tilde{T}_i, \delta_i)$ $i = 1, \cdots, n$ 로 나타낼 수 있다. 서로 다른 값을 갖는 관측된 자료를 순서대로 나열하면 $t_1 < t_2 < \cdots < t_d$ $(d \le n)$ 으로 주어지며 각 시점 t_i마다 다음과 같은 통계량을 표기할 수 있다. 즉, d_i 는 시점 t_i에서 관측된 사건 수이며 동점이 없으면 $d_i = 1$이 된다. Y_i는 시점 t_i- (즉, 시점 t_i의 바로 직전)까지 사건이 발생하지 않았거나 중도절단되지 않은 개체 수를 나타내는데 이를 위험에 처한 개체수라고 한다. 만약 시점 t_i 직전에 사건이 발생하였거나 중도절단된 개체들은 Y_i 에서 제외된다. 시점 t_i- 까지 생존하였다는 조건 하에서 t_i 에서 사건이 발생할 조건부 확률을 p_i 라고 하면 p_i 의 추정량은 $\hat{p}_i = \frac{d_i}{Y_i}$ 이며, 시점 t_i 에서의 조건부 생존확률은 $1 - \hat{p}_i = 1 - \frac{d_i}{Y_i}$로 추정된다. 카플란-마이어 추정량은 조건부 확률의 곱의 공식으로부터 구할 때 관측된 생존시간들 사이에서 발생하는 중도절단된 자료의 정보를 위험에 처한 개체 수 Y_i 에 반영하여 시점 t_i에서 다음과 같이 구하였다.

$$\begin{aligned} S(t_i) &= P(T > t_i) \\ &= P(T > t_i | T > t_{i-1}) P(T > t_{i-1} | T > t_{i-2}) \cdots P(T > t_1) \\ &= p_i \times p_{i-1} \times \cdots \times p_1. \\ \hat{S}(t_i) &= \prod_{j=1}^{i} (1 - \frac{d_j}{Y_j}). \end{aligned}$$

임의의 시점 t에서의 카플란-마이어 (KM) 추정량 $\hat{S}(t) = \prod_{t_i \le t} (1 - \frac{d_i}{Y_i})$가 되며 누적한계추정량 (product limit estimator)이라고도 한다. KM 추정량은 사건이 관측된 시점에서만 값이 감소하는 계단함수이며 감소하는 값은 사건발생 수와 중도절단 수에 의하여 정해진다. 그러나 마지막 관측값이 중도절단된 경우에는 KM 값이 더 이상 변하지 않으므로 관측이 종료된 시점 이후에도 생존함수의 추정이 정확하게 이루어지지 않게 된다. KM 추정량으로부터 누적위험률함수 추정량은 $\hat{H}(t) = -\log \hat{S}(t)$ 로 구해진다. 한편 Nelson (1972)은 누적위험률함수의 추정량을 구하여 생존함수를 추정하였다. Nelson-Aalen 누적위함률 추정량은 $\tilde{H}(t) = \sum_{t \le t_i} \frac{d_i}{Y_i}$이고, 생존함수 추정량을 $\tilde{S}(t) = \exp(-\tilde{H}(t))$이지만 이 추정량은 근사적으로 KM 추정량과 동일하다.

KM 추정량에 대한 분산 추정량은 Greenwood 공식 (Greenwood, 1926)에 의하여 구할 수 있으며 가우스과정으로 점근적으로 수렴한다는 것이 증명되어 있다 (Breslow와 Crowley, 1974). 이 외에도 KM 추정량은 생존함수의 비모수적 최대우도추정량이며 self-consistency를 만족한다 (Efron, 1967). 그러나 KM 추정량은 생존시간과 중도절단시간이 서로 독립이라는 가정을 전제하고 있으므로 이 독립성 가정이 위배되는 경우에는 편향적이고 왜곡된 추정량이 될 수 있다 (Fisher와 Kanarek, 1974; Scharfstein과 Robins, 2002). 한편, 1980년대에 들어서서 Aalen (1978)에 의하여 생존분석에서 다루는 추정량들을 계수과정 (counting process)으로 표현하게 되면서 점근적 분포에 대한 증명과정이 마팅게일 (martingale)의 중심극한정리를 이용하여 수월하게 이루어졌다. 따라서 KM 추정량의 점근적 성질도 계수과정으로 쉽게 증명되었다. 생존분석에 관한 확률과정이론에 대해서는 Fleming과 Harrinton (1990)과 Andersen 등 (1993)을 참조하기 바란다.

한편, 임상시험에서는 기존의 약과 신약의 효과를 비교할 때 두 그룹의 생존곡선이 유의하게 차이가 나는지를 검정하게 된다. 생존함수는 시간에 따라 값이 변하므로 관측된 시간의 전체 구간에서의 두 그룹의 생존율을 비교하는 방법을 고려해야 한다. 가장 많이 활용되는 검정법으로 로그-순의 검정법이 있으며 다음과 같이 주어진다.

$$\chi^2_{LR} = \frac{[\ \sum_{i=1}^{d}(d_{1i} - Y_{1i}\frac{d_i}{Y_i})\ ]^2}{\sum_{i=1}^{d}\frac{Y_{1i}Y_{2i}}{Y_i - 1}\frac{d_i}{Y_i}(1 - \frac{d_i}{Y_i})},$$

여기서 d_i, Y_i, d_{1i}, Y_{1i} 와 Y_{2i}는 시점 t_i $(i = 1, 2, \cdots, d)$ 에서 각각 사건 수, 위험에 처한 개체수, 그룹 1의 사건 수, 그룹 1의 위험에 처한 개체수와 그룹 2의 위험에 처한 개체수를 나타낸다. 위의 식에서 보는 바와 같이 로그-순위 검정통계량은 각 사건발생 시점에서의 기댓값과 관측값 간의 차이합을 표준화하여 제곱한 형태로 귀무가설 하에서 근사적으로 자유도가 1인 χ^2의 분포를 따른다. 로그-순위 검정법은 두 그룹의 생존함수가 비례할 경우에는 검정력이 높게 되는데 이는 사건발생 시점마다 관측값과 기댓값의 차이를 합하기 때문이다. 따라서 두 그룹의 생존함수가 비례하지 않거나 교차하게 되면 로그-순위 검정법의 합이 서로 상쇄가 되어 검정력이 약해진다. 로그-순위 검정통계량은 사건 발생시간의 순위에 점수를 준 형태이므로 점수 대신에 가중함수를 주어 선형순위검정법으로 일반화하여 시간 구간 중 중도절단의 분포에 따라 가중치를 다르게 주어 검정력을 높일 수 있다 (Klein Moschberger, 2010, 7장). 만약 두 그룹의 생존함수가 비례적인 가정에서 크게 위배되는 경우에는 두 생존함수의 차이가 최대가 되는 시점에서의 차이값을 이용하는 콜모고로프 검정법이나 중앙값의 차이에 기반한 Median 검정법 (Chen Zhang, 2016)을 이용하는 것이 더 검정력이 높다.

3.1. 칵스 회귀모형

의학, 보건학 및 역학분야에서는 질환과 관련하여 나이, 성별, 인종 등과 같은 인구통계학적 요인과 비만도, 혈압, 인슐린 수치, 체내 지방량 등과 같은 임상학적 요인 및 흡연과 음주 습관, 환경오염도 등과 같은 환경학적 요인들을 중심으로 생존시간에 유의한 위험요인 (risk factors)들을 찾아내기 위하여 수많은 연구들이 진행되었다. 특히 임상시험연구에서 생존시간에 유의한 영향을 주는 요인들을 찾아내어 환자들의 생존율을 향상시키는 신약을 개발하거나 식이습관이나 환경적 요인을 변화시켜 국민건강을 위한 정책들을 수립하는데 생존분석에서 구축된 통계예측모형들이 중요한 역할을 하였다. 생존분석에서 가장 크게 활용되는 회귀모형은 위험률함수의 비에 대하여 예측요인들의 선형모형을 가정한 칵스모형 (Cox, 1972)으로 다음과 같이 주어진다.

$$h(t|\mathbf{x}) \quad = \quad h_0(t)\exp(\mathbf{x}\beta),$$

여기서 $\mathbf{x}$는 $p \times 1$ 벡터로 p 개의 공변량을 나타내며 β는 이 공변량에 대응되는 $1 \times p$ 회귀계수벡터이다. $h(t|\mathbf{x})$ 는 공변량 $\mathbf{x}$ 가 주어져있을 때 위험률함수이며 $h_0(t)$는 공변량이 $\mathbf{x} = 0$ 일 때의 기저위험률함수 (baseline hazard function)로 이 모형에서 특정한 분포가 주어지지 않는다. 이 식의 양변에 로그를 취하게 되면 다음과 같이 주어진다.

$$\log\frac{h(t|\mathbf{x})}{h_0(t)} \quad = \quad \mathbf{x}\beta.$$

칵스모형은 위험률함수의 로그비가 예측인자의 선형회귀식으로 표현되며 회귀계수 β는 상대위험률로 해석될 수 있다. 위의 식으로부터 서로 다른 공변량의 값을 갖는 개체들의 위험률의 비가 시간과 상관

없이 일정하기 때문에 이 모형을 비례위험모형 (proportional hazards model)이라고도 한다. 비례위험 가정은 위험률의 비가 시간과 독립적이라는 매우 강한 조건이지만 비교적 짧은 구간에서는 이 모형이 잘 적합된다. 칵스모형은 준모수적인 (semi-parametric) 모형이라고 할 수 있는데 이는 기저위험률함수에 대하여 어떠한 분포도 가정하지 않고 회귀계수에 대한 모형만 가정하기 때문이다. 실제로 위험요인이 위험률함수에 미치는 효과를 나타내는 회귀계수에 대한 추정을 위하여 Cox (1975)는 부분가능도함수 (partial likelihood function)를 제안하였는데 이 식에서 기저위험률함수는 전혀 고려하지 않는다.

$$PL(\beta) \quad = \quad \prod_{i=1}^{n} \frac{\exp(\mathbf{x_i}\beta)}{\sum_{l \in R(t_i)} \exp(\mathbf{x}_l\beta)},$$

여기서 $R(t_i)$은 시점 t_i 에서의 위험에 노출된 개체집합으로 t_i 이전에 사건이 발생하지 않았거나 중도절단되지 않은 개체들을 나타낸다. 그러나 생존함수를 추정하기 위해서는 기저위험함수의 추정량이 필요하므로 Breslow (1975)의 비모수적 추정량을 이용한다. 한편, 칵스모형에 대한 적합성을 진단하기 위하여 회귀진단에서와 같이 잔차를 이용한 통계량들이 많이 제안되어있다. 예를 들어, 칵스-스넬 잔차 (Cox와 Snell, 1968), 마팅게일 잔차 (martingale residual)와 이탈도 잔차 (deviance residual) 등을 이용하여 그림으로 적합성을 진단하거나 검정통계량으로 적합성검정을 할 수 있다. 일반적으로 비교적 짧은 구간에서는 칵스모형이 잘 적합되지만 (Lee와 Lee, 2017) 오랜 기간동안 추적조사하는 경우에는 칵스모형의 비례위험가정이 성립하지 않는 경우가 종종 발생하여 칵스모형의 비례위험가정을 일반화하여 층화칵스모형 (stratified Cox model)이나 시간종속 칵스모형 (time-dependent Cox model) 등으로 확장시키는 연구결과들이 많이 있다 (Klein과 Moeschberger, 2010). 한편, Aalen (1989, 1993)은 다음과 같이 칵스모형과 달리 위험률함수를 기반으로 시간종속 독립변수와 시간종속 회귀계수를 고려한 비모수적이고 가법적인 (additive) 회귀모형을 제안하였다.

$$h(t|\mathbf{x_i}(t)) \quad = \quad \beta_0(t) + \sum_{k=1}^{p} \mathbf{x}_{ik}(t)\beta_k(t).$$

Aalen이 제안한 위의 모형은 칵스모형에 비하여 모형에 대한 가정이 약하여 일반적인 경우로 크게 활용될 수 있지만 통계적 추론과정에서 시간의 함수인 회귀계수에 대한 직접접인 추정과정이 어렵고 불안정적이다. 대신에 시점 t까지 회귀계수를 누적하여 누적된 위험함수값을 추정하고 이 추정량의 분산을 추정하여 통계적인 추론을 할 수 있다. 이러한 복잡한 추론과정을 극복하기 위하여 시간에 종속된 회귀계수의 가정을 상수로 단순화한 모형을 Lin과 Ying (1994)이 제안하였다.

3.2. 모수적 선형회귀모형

생존분석에서도 기존의 회귀모형과 같이 예측요인들을 독립변수로 하고 생존시간을 종속변수로 간주하는 선형회귀모형을 적합할 수 있다. 이 경우 생존시간이 음수가 아닌 실수 공간의 범위에서 관측되므로 로그를 취하여 로그생존시간을 종속변수로 간주하여 실수 전구간에서의 선형모형을 적합시킨다. 또한 생존시간은 정규분포와 같이 대칭분포를 이루는 경우가 거의 없고 대부분 비대칭적이며 치우친 분포를 띠고 있으므로 로그변환을 하는 것이 추론하기에 더 적합하다. 따라서 종종 이 모형을 로그선형회귀모형이라고 하며 다음과 같이 주어진다.

$$\log T \quad = \quad \mu + \mathbf{x}\beta + \sigma\epsilon.$$

위에서 μ와 σ는 각각 공변량이 없는 경우의 평균벡터와 척도모수를 나타내며 ϵ은 오차항을 나타내는 확률변수로 이 오차항의 분포에 모수적인 분포를 가정하게 된다. 주로 활용되는 모수적 분포는 생존

분석에 적합한 지수분포, 와이블분포, 감마분포, 로그-정규분포, 로지스틱 분포 등이 있다. 로그선형회귀모형에서 $S(t|\mathbf{x}) = S_0(t\exp(-\mathbf{x}\beta))$으로 $S_0(t)$는 공변량이 $\mathbf{x} = 0$일 때의 기저생존함수를 나타낸다. 이 관계에서 $\exp(-\mathbf{x}\beta)$를 생존율을 가속화하는 요인 (accelaration factor)라고 하는데 이는 기저생존함수에서의 시간을 공변량의 변화값을 곱하여 가속화하는 효과를 가지게 되기 때문이다. 예를 들어 어떤 공변량의 값을 갖는 개체의 어느 시점에서의 생존율은 기저생존함수에서 가속화요인의 배만큼 곱한 이후의 시점에서의 생존율과 동일하다는 것으로 해석된다. 따라서 이 모형을 가속수명모형(accelerated failure time model) 이라고도 한다. 만약, β 값이 양수이면 시점 t에서의 생존율은 가속화요인의 배만큼 감소된 시점에서의 기저생존율과 동일하게 되기 때문에 결과적으로 생존시간이 증가되는 효과가 있다. 앞 절에서 소개한 칵스모형에서의 회귀계수와 부호가 반대가 되는 결과를 갖게 된다. 모수적 선형회귀모형에서 회귀계수와 오차항의 척도모수를 추정하는 방법은 모수적으로 가정한 분포를 이용하여 가능도함수 (likelihood function)를 구하고 이 가능도함수를 최대화하는 최대가능도추정량 (maximum likelihood estimator)을 취한다. 예를 들어, 지수분포를 따르는 생존자료의 로그선형 회귀모형에서의 가능도함수는 다음과 같이 주어진다.

$$L(\beta) = \prod_{i=1}^{n}(h_i(t|\mathbf{x_i}))^{\delta_i} S(t|\mathbf{x_i}) = \prod_{i=1}^{n}\exp(\mathbf{x_i}\beta)^{\delta_i}\exp(-\exp(\mathbf{x_i}\beta)t_i),$$

여기서 δ_i가 중도절단여부를 나타내며 가능도함수는 위험률함수와 생존함수의 곱으로 표현된다. 최대가능도추정량과 이 추정량의 분산을 구하기 위해서 가능도함수에 로그를 취하여 한 번 미분한 스코어함수 (score function)와 두 번 미분한 정보행렬 (information matrix)을 구하면 다음과 같이 주어진다.

$$\begin{aligned} U_k(\beta) &= \frac{\partial l(\beta)}{\partial \beta_k} \\ &= \sum_{i=1}^{n} x_{ik}(\delta_i - t_i\exp(\mathbf{x_i}\beta)), \quad k = 1, 2, \cdots, p, \\ I(\beta_k, \beta_l) &= -(\frac{\partial^2 l(\beta)}{\partial\beta_k\partial\beta_l}) \\ &= \sum_{i=1}^{n} x_{ik}x_{il}t_i\exp(\mathbf{x_i}\beta), \quad k, l = 1, 2, \cdots, p. \end{aligned}$$

위의 스코어함수의 해가 최대가능도추정량이며 이 해를 구하기 위하여 수치해석방법의 하나인 Newton-Raphson 기법을 사용한다. 또한 최대가능도추정량의 분산추정량은 정보행렬로부터 구해진다.

모수적 회귀모형에서도 모수적 분포가정이 적합한지 여부를 진단하고 이상점과 같은 자료에 대하여 회귀진단을 하는데 이 경우에는 표준화잔차를 이용하거나 지수분포나 와이블분포에서는 칵스-스넬 잔차나 마팅게일잔차를 이용한다.

4. 유전체 자료분석을 위한 생존분석방법

인간게놈 프로젝트가 수행된 후 대용량의 유전체 자료가 수집되면서 유전정보가 생존시간에 미치는 영향에 대한 연구가 활발하게 이루어지기 시작하였다. 그러나 표본 수에 비하여 유전정보를 나타내는 독립변수의 수가 엄청나게 많기 때문에 기존의 통계적 방법으로는 모든 유전정보를 동시에 고려하는 모형을 적합할 수 없으므로 우선적으로 유전체 자료를 하나씩 적합하여 유의한 유전자를 찾아내는 접근방법을 시도하였다. 이 접근방법은 다중검정법 (multiple testing) 문제를 야기하게 되어 가장 용이하게

적용할 수 있는 집단별 오류율을 조절하는 Bonferroni 방법을 적용하였다. 그러나 유전정보의 수가 증가할수록 Bonferroni 방법은 너무 보수적으로 제 1종 오류를 조절하게 되어 유의한 유전자를 찾아내는 것이 쉽지 않았다. 이에 대한 대안책으로 위발견율 (false discovery rate; FDR)이라는 새로운 형태의 오류율을 Benjamini과 Hochberg (1995)가 제안하였다. FDR로 제 1종 오류를 조절할 경우 Bonferroni 방법보다 덜 보수적인 기준을 이용하므로 유의한 유전자 정보를 더 많이 식별할 수 있게 된다.

한편, 일반적으로 유전체 자료들은 서로 독립적이지 않고 연관성 있게 작동하기 때문에 하나씩 적합된 단순회귀모형이 아니라 동시에 적합된 다중회귀모형을 고려해야 한다. 그러나 이 경우 적합하고자 하는 독립변수의 수가 표본 수보다 훨씬 많아 기존의 회귀모형에서 회귀계수에 대하여 제약조건으로 벌점(penalty)를 주는 통계적인 방법들이 기존의 칵스 모형으로 확장되어 개발되었다. 이 절에서는 다양한 벌점함수를 활용한 칵스 모형과 칵스부스트 모형에 대해서 살펴보고자 한다.

4.1. 벌점화 칵스모형

의학이나 임상학 분야에서 환자들의 유전정보와 질환과의 연관성에 대한 초기 연구결과는 임상학적으로 동일한 진단을 받은 환자들의 생존시간이 마이크로어레이에 의해 밝혀진 유전정보에 따라 유의하게 달라진다는 것이었다. 환자들의 유전자의 차이로 인해 임상학적으로 동일한 그룹이었지만 생존시간이 유의하게 다르다는 분석결과가 발표되면서 유전체 자료를 고려하는 통계적 모형을 개발할 필요성이 증폭되었다 (Golub 등, 1999; Alizadeh 등, 2000; Garber 등, 2001; Beer 등, 2002). 초기의 마이크로어레이 자료는 수천 개에 불과하였지만 현재 수집되는 자료는 거의 수만 개에 이르고, SNP 칩에서 한꺼번에 수집되는 SNP 수는 백만 개를 넘는다. 이 외에도 RNA-seq, protein arrays, whole genome sequencing 자료들을 동시에 고려하게 되면 회귀모형에서 다루는 예측인자의 차원이 천문학적인 수가 된다. 이 중에서 질환이나 생존시간에 실제로 유의한 영향을 미치는 유전자를 찾기 위하여 회귀모형에 적합하려면 표본 수보다 엄청나게 많은 회귀계수에 대한 벌점함수 (penalty function)를 고려하여 추정하는 방법이다. 이러한 이론적 근거는 이와 같이 천문학적으로 많은 유전자 중에서 실제로 질환과 연관된 유전자의 수는 극히 적다라는 희박성 (sparsity) 에 대한 가정이다 (Hastie 등, 2011). 대표적인 벌점함수는 Lasso, Ridge, Elastic-Net 이 있으며, 이 절에서는 칵스모형을 중심으로 벌점화하는 방법에 대하여 살펴보겠다.

4.1.1. Lasso-Cox, adaptive Lasso-Cox, fused Lasso

Tibshirani (1997)는 생존분석의 칵스모형에 l_1-norm 제약조건 하에서 음의 로그부분가능도함수를 최소화하는 Lasso (least absolute shirinkage and selection operator) 벌점화 추정량을 제안하였다.

$$\hat{\beta}_{Lasso} \quad = \quad \text{argmin}[\ -\sum_{i=1}^{n}\delta_i(\mathbf{x_i}\beta - \log(\sum_{j=1}^{n}\exp(\mathbf{x_j}\beta))) + \lambda\sum_{k=1}^{p}|\beta_k|\ ].$$

위의 식에서 $\lambda = 0$인 경우는 제약조건이 없는 칵스모형에서 구한 추정량과 동일하며 $\lambda \to \infty$로 커질수록 희박성 (sparsity) 조건이 강화되어 대부분의 회귀계수의 값이 정확하게 0의 값을 취한다. 따라서 λ의 값에 따라 희박성 조건이 조절되는 효과가 있으므로 이를 조율모수 (tunning parameter)라고 한다. 최적의 λ는 일반화된 교차타당성 (generalized cross-validation) 통계량 (Wahba, 1980)을 최소화하는 값으로 구한다. Lasso는 고차원자료로부터 희박성의 가정이 만족하는 조건 하에서 적은 수의 유의한 정보를 갖는 변수를 선택하는 방법이며 회귀계수의 값이 0가 아닌 최대변수의 수는 표본의 수 n를 넘지 않는다.

Lasso가 모든 β를 크기에 상관없이 동일하게 간주하고 있는 점을 개선하기 위하여 β의 크기에 따라 다른 가중치를 주는 방법을 제안하였는데 이 방법을 Adaptive Lasso-Cox 방법이라고 한다 (Zhang과

Lu, 2007). Adaptive Lasso-Cox 방법은 벌점함수를 $\lambda \sum_{k=1}^{p} |\beta_k| \tau_k$로 주는데 여기서 각각의 β_k에 다른 τ_k의 가중치를 주어 β의 크기에 대한 정보를 조절한다. 다시 말하면 큰 β_k에는 작은 τ_k 값을, 작은 β_k에는 큰 τ_k를 주어 효과가 큰 변수에는 작은 벌점을 주고 효과가 작은 변수에는 큰 벌점을 줌으로써 유의미한 변수를 더 잘 선택하도록 해준다.

만약, 고차원의 예측인자간에 순서의 개념이 있는 경우에는 각 변수의 절대값에 대한 벌점을 주는 Lasso 방법을 개선하여 연속적인 두 변수의 차이의 절대값에 대한 벌점도 추가적으로 고려하여 희박성의 조건을 더 효율적으로 활용하는 Fused Lasso 방법을 Tibshirani 등 (2005)이 제안하였다. Fused Lasso 의 벌점함수는 $\lambda_1 \sum_{k=1}^{p} |\beta_k| + \lambda_2 \sum_{k=2}^{p} |\beta_k - \beta_{k-1}|$로 주어진다. 이 방법은 예측인자인 유전체 자료에 순서의 개념이 없는 경우에도 heat map을 이용한 군집분석 결과로부터 유전체 자료에 순서를 매겨 응용할 수 있다. 또는 특정한 유전자와 근거리에 있는 유전자들에 한해서 적용할 수도 있다.

4.1.2. Ridge-Cox

Ridge 회귀모형은 Hoerl과 Kennard (1988)이 제안한 것으로 다중선형회귀모형에서 상관성이 높은 독립변수들로 인하여 발생하는 다중공선성 (muti-collinearity) 문제를 해결하기 위한 방법이었다. 다중공선성은 회귀모수의 추정량의 분산을 크게 증가시켜 추정량의 신뢰도를 낮게 하기 때문에 회귀계수의 제곱값에 제한조건을 주어 분산의 증가를 줄여 안정적인 추정량을 제공하였다. Ridge의 벌점함수는 l_2-norm이며 다음과 같이 주어진다.

$$\hat{\beta}_{Ridge} = \text{argmin}[\ -\sum_{i=1}^{n} \delta_i (\mathbf{x_i}\beta - \log(\sum_{j=1}^{n} \exp(\mathbf{x_j}\beta))) + \lambda \sum_{k=1}^{p} {\beta_k}^2\ ].$$

Ridge 방법은 변수들 간의 상관성을 고려하기 때문에 고차원의 유전체 자료들을 분석하는데 적합하지만 변수선택하는 과정에서 모든 변수들이 0이 아닌 계수를 가지므로 Lasso와 달리 변수선택이 가능하지 않다. 따라서 변수가 매우 많은 경우에는 Lasso 방법으로 비교적 적은 수의 변수를 선택하도록 한다. 그러나 adaptive Lasso-Cox 방법에서 β에 주는 가중치를 추정하기 위하여 Ridge 방법으로 구한 β의 값의 역수를 사용하는데 활용될 수 있다.

4.1.3. Elastic-Net, elastic Net-Cox

Elastic-Net (EN) 방법은 Lasso에서의 희박성과 Ridge의 상관성의 장점을 취하여 이 두 방법의 벌점함수를 가중치로 주는 방법이다. 일반적인 회귀모형에 대하여 Zou와 Hastie (2005)이 EN 방법을 처음 제안하였는데 Simon 등 (2011)이 EN 방법을 칵스모형으로 확장하여 EN-Cox 방법을 제안하였다. EN-Cox 방법에서의 제약조건은 다음과 같이 주어진다.

$$\hat{\beta} = \text{argmin}[\ -\sum_{i=1}^{n} \delta_i (\mathbf{x_i}\beta - \log(\sum_{j=1}^{n} \exp(\mathbf{x_j}\beta))) + \lambda(\alpha \sum_{k=1}^{p} |\beta_k| + (1-\alpha) \sum_{k=1}^{p} {\beta_k}^2)\ ].$$

EN-Cox 방법의 벌점함수는 Lasso의 l_1-norm과 Ridge의 l_2-norm을 조합하여 두 방법의 장점을 가지게 되어 Lasso와 달리 표본의 수보다 더 많은 변수도 선택할 수 있으며 변수들 간의 상관성도 고려할 수 있다. EN-Cox 방법에서는 회귀계수를 추정하기 위하여 cyclic coordinate descent 알고리듬을 개발하여 실제 자료분석에서 계산시간을 기존의 알고리듬 비하여 열 배 (54.9 vs. 679.5 (초)) 가깝게 줄이는 결과를 보여 주었다 (Simon 등, 2011).

지금까지 살펴본 방법들을 실제 자료분석에 적용하여 비교한 결과를 Zou와 Hastie (2005)는 (Table 4.1)로 정리하였다. 이 결과는 전립선암 (Stamey 등, 1989) 환자들의 자료에서 전립선항원촉매 반응에

대하여 8개의 예측변수들을 회귀모형으로 적합할 경우 OLS (최소제곱법), Ridge, Lasso, Elastic-Net 방법을 적용하여 각 방법의 시험군의 평균제곱오차와 선택된 변수들을 나타낸 것이다.

Table 4.1 Prostate cancer data: comparing different methods

Method	Test mean-squares error	Variable selected
OLS	0.586(0.184)	All
Ridge	0.566(0.188)	All
Lasso	0.499(0.161)	(1,2,4,5,8)
Elastic Net	0.381(0.105)	(1,2,5,6,8)

위의 표의 결과에서 OLS에 비하여 Ridge 방법은 8개의 변수를 모두 선택하지만 Ridge 방법에서 예측오차가 약간 줄어든 결과를 보여준다. 그러나 이 두 방법에 비하여 Lasso와 EN 방법은 8개의 변수 중 5개를 선택하면서 예측오차도 작아지는 결과를 보여주고 있다. 특히 EN 방법은 예측오차가 다른 3가지 방법에 비하여 매우 작아지는 것을 알 수 있다.

4.2. 칵스 부스트

고차원 자료를 분석하여 통계적 모형을 구축할 때 활용되는 부스팅 (boosting) 기법은 기계학습방법에서 활용되는 앙상블 방법으로 다수의 예측모형을 생성한 후 가중치를 사용하여 합치는 기법이다. Freund와 Schapire (1995)가 제안한 AdaBoost (adaptive boost) 알고리듬은 통계적 추정과 모형을 구축하는 과정에서 gradient descent 알고리듬을 응용하여 단계적이고 가법적인 모형화 (stagewise and additive modelling)방법으로 이론화되었다 (Friedman 등, 2000). 부스팅 방법의 기본적인 개념은 다수의 예측모형을 연속적으로 재가중치된 자료로부터 반복적으로 생성하여 최종적으로 다수의 예측모형을 선형결합으로 취하는 방법이다. 만약 반복회수가 $m = 1, \cdots, M$ 일 때 m 단계에서 재가중치된 자료는 $m-1$ 단계에서 구한 예측모형에만 의존하며 그 전 단계의 $m-2, m-3, \cdots$ 의 결과에는 독립적이다. 부스팅은 하나의 예측모형보다는 다수의 예측모형을 결합하여 평균을 내면 예측율이 더 향상된다는 원리에 기반하며 Brieman (1998)에 의하여 functional gradient descent 알고리듬으로 해석되었다. 이후에 Friedman 등 (2000) 에 의하여 부스팅은 함수추정방법으로 재해석되면서 이진형 반응변수의 분류문제, 연속형 반응변수의 회귀모형과 생존분석의 칵스모형의 예측력을 높이는 방법으로 활용되었다. 부스팅의 알고리듬에 대한 자세한 이론적 개념과 설명은 Buhlmann과 Hothorn (2007)을 참조하라.

생존분석에서 생존시간과 연관성이 있는 유전자 정보를 고려하는 고차원의 유전체 자료들을 기반으로 통계예측모형을 구축할 때 기존의 임상적인 위험인자들이 필수적으로 포함되어야 한다. 이러한 임상적인 위험인자들을 기반으로 정의된 예후지수 (prognostic index)는 유전자 정보와는 별도로 환자들의 생존시간에 유의한 영향을 주는 것으로 변수선택과정에서 반드시 포함되도록 하고 이 변수들의 추정량이 어떠한 제약조건에 의해서도 영향을 받지 않아야 한다. 그러나 지금까지 살펴본 벌점함수를 이용하는 방법들은 예후지수와 관련된 위험인자들에 대하여 별도로 제약조건에서 제외시키지 않고 있다. 다시 말하면 수많은 유전자 요인들과 임상적인 요인들을 동시에 고려하여 변수선택을 할 때 모든 변수에 벌점함수를 동시에 주기 때문에 필수적으로 고려해야 할 변수들과 그렇지 않은 변수들을 분리하여 추정할 수 없었다. 이러한 문제점을 해결하기 위하여 Binder와 Schmacher (2008)는 칵스모형에 대하여 부스팅 기법을 확장한 칵스 부스트 (Cox boost) 방법을 제안하였다. 이 방법은 칵스모형에 부스팅 기법을 활용할 때 부스팅을 하는 단계마다 offset-based gradient boosting 알고리듬을 적용하여 그 다음 단계의 모형을 적합하는 것이다. 필수적으로 포함해야 할 변수들을 offset으로 간주하여 이 변수들에 대해서는 벌점화하지 않고 그 외 변수들에 대해서만 벌점함수를 적용하여 추정량을 update하는 것이다.

이 방법으로 예측력이 얼마나 향상되는지 비교하기 위하여 Diffuse large B-cell lymphoma 환자의 실제자료를 분석하였다. 이 자료는 240명의 환자로 구성되어 있으며 추적기간의 중앙값은 2.8 년이며 7399개의 마이크로어레이 자료가 관측되어 있다. 또한 이 중 220명의 환자들은 임상자료로부터 계산된 IPI (international prognostic index) 점수에 대한 정보를 가지고 있었다. IPI 점수가 220명의 환자에게만 얻어졌으므로 220명의 환자자료를 이용하여 다음의 4가지 다른 모형을 적합하였다: 1) IPI 점수만 고려한 칵스모형 2) 마이크로어레이 자료만 적합한 칵스부스트 모형 3) IPI와 마이크로어레이자료를 동시에 고려하면서 IPI의 포함여부를 옵션으로 고려한 모형 4) IPI와 마이크로어레이자료를 동시에 고려하면서 IPI의 포함여부를 필수적으로 고려한 모형. 위의 4가지 모형을 적합한 후 Kaplan-Meier 곡선 (Binder와 Schmacher, 2008)과 각각의 예측모형에서 추정된 생존곡선의 예측오차를 비교한 결과, IPI와 마이크로어레이 자료를 각각 포함하는 모형보다는 동시에 고려하는 모형의 예측오차가 적었으며 특히 IPI 점수를 옵션으로 간주하는 모형보다는 필수적으로 포함하는 모형의 예측오차가 적었다.

결론적으로 칵스부스트는 임상적인 자료와 고차원의 유전체 자료를 동시에 고려하면서 융통성있게 벌점화하는 방법을 제공하였으며 실제적으로 추정하는 과정에서 벌점화하지 않는 임상자료의 추정방법을 효율적으로 하기 위하여 offset-based 부스팅 기법을 제안하였다. 또한 실제자료분석 결과를 통하여 임상적 인자들을 포함하는 경우과 그렇지 않은 경우에 마이크로어레이 자료의 효과의 크기 순서가 바뀌게 되는 경우가 발생하기도 하고, 반대로 마이크로어레이 자료를 동시에 고려하게 되면 임상적인 위험인자의 효과에도 영향을 준다는 것을 알 수 있었다. 이는 유전체 자료에 대한 생존분석방법에서 기존 연구결과로부터 나온 임상적 위험인자들의 효과도 동시에 고려해 주는 것이 필요하다는 것을 암시하고 있다. 실제적으로 칵스부스트를 수행하는 소프트웨어는 (`https://cran.rproject.org/web/packages/CoxBoost/`)를 참조하라.

5. 기계학습을 활용한 생존분석방법

정보화시대를 거쳐 엄청나게 빠른 속도로 대용량의 자료들이 누적되면서 사회 각 분야에는 빅데이터를 분석할 수 있는 방법들에 대한 연구들이 활발하게 이루어지고 있다. 빅데이터의 특징은 비정형적이고 복잡하며 다양한 형태의 자료들로 대용량으로 수집된 자료라는 것이다. 기계학습방법은 이와 같이 복잡하고 비선형적인 빅데이터로부터 컴퓨터가 학습을 통하여 일정한 패턴을 유추하고 체계적으로 분류하여 예측오차를 줄이기 위해서 많은 통계적인 이론과 확률론적인 원리를 활용하는 인공지능 기법이다. 예를 들면, 우편국에서 손글씨로 적힌 주소를 컴퓨터에서 자동인식하거나 이미지분석을 통하여 개체를 식별하거나, 개인의 신용카드 사용내역을 통하여 리스크 분석을 하는 등 수많은 자료들로부터 컴퓨터가 반복적인 학습을 통하여 유용한 정보를 추출하여 통계적 추론을 하는 것이다. 기계학습방법은 주로 분류분석 (classification) 문제를 다루는데 매우 유용하게 활용되고 있다, 생존분석에서도 암질환을 진단하거나 환자의 예후 및 치료방법의 결정과 생존여부를 예측하는 경우에 분류분석을 해야하는 경우가 흔히 발생한다. 기존의 생존분석 방법으로 예측모형을 구축할 수도 있으나 대용량의 유전체 정보를 활용하여 분류분석을 하기 위해서는 자연스럽게 기계학습방법을 접목하여 활용할 수 있다. 이미 중도절단자료가 많은 생존분석에서도 기계학습방법은 이미 이십여 년전부터 암질환 진단과 재발여부를 예측하는 기법으로 활용되어 왔다 (Cruz와 Wishart, 2006). 이 절에서는 기계학습방법이 생존분석에 활용된 통계적인 방법들 중에서 생존나무 (survival tree), 인공신경망 (artificial neural network), support vector machine과 앙상블 (ensemble)에 대하여 살펴봄으로써 한 개인으로부터 여러 형태로 관측되는 유전체자료들을 융합하여 생존시간을 예측하는 모형을 발전해 나가는데 기초지식을 접하는 계기를 제공하고자 한다.

5.1. 생존나무

Tree-based 방법은 분류작업이나 예측하는 기준을 이해하기에 직관적이고 해석이 용이하여 여러 분야에서 응용되어져 왔다. 응용통계분야에서도 CART (classification and regression tree) 방법 Breiman 등 (1984)에 의해 제안된 이후에 크게 활용되고 있는 기법이다. 중도절단자료를 다루는 생존분석에서도 생존나무 (survival tree) 방법이 개발되어 많은 임상의들이 환자들을 진단하는 방법으로 활용되고 있다 (Yoo, 2018). 일반적인 결정나무 (decision-tree)방법과 생존나무 방법의 차이점은 분류하는 기준이 다르다는 것인데, 결정나무의 경우 분류기준이 각 변수의 값에 의해서 결정되는 반면에 생존나무에서는 서로 다른 마디 (node) 간의 이질성 (heterogeniety)을 높여주거나 동일한 마디 내의 동질성 (homogeniety)을 높여주는 분류기준을 사용한다는 것이다. 마디 내의 동질성을 측도하는 손실함수를 최소화 하기 위하여 거리의 개념을 사용하는 방법도 제안하였지만 Ciampi 등 (1986)의 논문에서는 마디 간의 이질성을 검정하는 로그-순위 검정통계량을 분류기준으로 제안하였다. 그 외에도 서로 다른 마디 간의 이질성을 측도하는 통계량으로 가능도비를 적용하거나 가중선형검정법을 제안하였다. 한편 생존나무를 구축하는데 있어서 최종나무를 결정하기 위해서 가지치기 방법과 나무크기를 정하는 문제가 중요하다. LeBlanc와 Crowley (1993)의 논문에 이 문제에 대한 이론적인 결과와 알고리듬이 잘 설명되어 있으며 폐암의 임상적인 실제자료분석을 통하여 생존나무 방법을 적용한 결과 환자군을 4 그룹으로 분류하여 그룹 간의 카플란-마이어 곡선이 유의하게 달라지는 것을 보여주었다 (LeBlanc와 Crowley, 1993). 이와 같이 생존나무 방법을 이용하여 환자들의 예후나 치료법에 따라 생존율이 달라지는 것을 직관적으로 이해할 수 있게 되고 진단하는데 도움을 줄 수 있다. 생존나무를 수행하는 소프트웨어는 (`https://cran.rproject.org/web/packages/rpart/`)를 참조하라.

5.2. 인공신경망

암 질환을 연구하는 분야에서는 인공신경망 (artificial neural network; ANN) 방법은 1980 년대부터 암 질환을 식별하고 진단하는 데 유용하게 쓰여진 기법이다. 인공신경망 방법은 통계학에서 주로 다루는 선형적인 함수가 아닌 비선형적이고 상호작용을 고려하는 복잡한 함수를 사용하여 여러 개의 은닉 층 (hidden layers)에서 가중치를 순차적으로 추정하여 결과변수를 산출한다. 인공신경망 방법을 수행하기 위해서는 은닉 층의 개수, 결과변수를 산출해 내는 활동함수 (activation function), 훈련군 알고리듬 및 과적합 (over-fitting)을 피하기 위한 종료시점 등 많은 변수들을 정해야 한다. 이미 임상분야에서 많은 연구결과들이 ANN 방법에 의해서 분석되어졌는데 마이크로어레이자료를 기반으로 양성 (benign)과 악성 (malignancy)을 분류하는 것부터 시작하여 암 질환의 재발과 생존여부를 예측하는 방법으로 광범위하게 활용되어져 왔다. 예로써, Ayer 등 (2010) 연구에서는 18, 269명의 유방암 환자들로부터 얻어진 48, 744의 mammographic imaging 자료와 인구통계학적인 요인 및 임상학적인 요인을 기반으로 1000개의 은닉 층 마디를 갖는 3층 피드포워딩 (feed forwarding) 인공신경망을 구축하였으며 과적합을 피하기 위하여 early stopping 기법을 적용하였다. 이 ANN을 10-fold-cross-validation 한 결과 모형의 예측력을 평가하는 지표 중의 하나인 AUC (area under the receiver operating characteristic curve)의 값을 0.965로 얻었다. 이 결과는 진단임상의로부터 얻은 AUC=0.939 보다 더 높은 수치이었으며 ANN으로부터 더 좋은 결과를 얻을 수 있다는 것을 의미한다. 또한 이 AUC의 calibration plot도 살펴본 결과 정확도가 매우 높게 나왔다. 이 외에도 임상분야에서 스크리닝 검사에서 얻어지는 수많은 분자생물학적 자료들과 임상학적 요인들을 입력하여 인공신경망 기법에서 요구되는 은닉 층과 통계적 함수를 설정하여 암 질환 여부를 예측하는 분석결과들이 쏟아져 나오고 있다. 그러나 인공신경망은 입력된 자료로부터 결과변수가 나오는 과정이 'Black Box'와 같이 어떠한 관계식으로부터 나오는지 해석하는 것이 어렵다는 단점이 있다. 인공신경망을 수행하는 소프트웨어는 (`https://cran.rproject.org/web/`

packages/neuralnet/)를 참조하라.

5.3. Support Vector Machine

Support vector machine (SVM) 방법은 이진형 반응변수에 대하여 두 그룹 간의 경계 (margin)을 최대화시키며 오분류율을 최소화하는 초평면 분리자 (separating hyper-plane)를 찾는 분류방법으로 이 초평면 분리자에 가까운 자료들을 support vector (SV)라고 명명한다. 분리자를 찾는 과정에서 SV와의 거리를 최대화시키는 데 비선형적인 경우에는 원자료에 고차원의 커널함수 (kernel function)를 이용하여 고차원에서 선형적으로 분리될 수 있도록 하여 비선형적이고 고차원적인 자료에 대해서도 쉽게 분류할 수 있게 한다. SVM 방법은 연속형 반응변수에 적합한 회귀모형으로 확장되었고 중도절단자료를 다루는 생존분석에 대해서는 Shivaswamy 등 (2007)에 처음으로 제안되어 벌점의 오류에 대한 마진은 개선하였으나 가중치에 대하여 고려하지 않았다. 이후 Khan과 Zubek (2008)이 이러한 제한점을 개선하는 SVRc (support vector regression for censored data)를 제안하였다. 이 방법은 비대칭적인 손실함수를 이용하여 생존시간이 관측된 경우와 중도절단된 경우에 벌점의 가중치를 다르게 조절하는 방법으로 오분류율을 최소화하였다. 임상에서의 실제자료에 대하여 각스모형과 비교한 결과 SVRc의 방법에서 예측결과가 더 높게 나오는 것을 보여주었고 이 방법으로 분류된 고위험군과 저위험군의 위험률비가 더 높게 나와 두 위험군의 차이를 극대화하는 분류방법이라는 것을 보여 주었다. SVM을 수행하기 위해서는 소프트웨어 (https://cran.rproject.org/web/packages/e1071/)를 참조하라.

5.4. 앙상블

앙상블 (ensemble) 방법은 다수의 분류자를 생성하여 각각의 분류자의 정확도를 고려하여 가중치를 주어 새로운 하나의 분류자를 만드는 방법이다. 이 방법은 하나의 분류자보다 다수의 분류자가 더 예측력이 높아질 수 있으며 가중치를 적절하게 조절하면 예측력이 낮은 분류자를 보완할 수 있다는 개념에서 도출되었다. 앙상블의 대표적인 방법으로 bagging, random forest, boosting이 있으며 다수의 방법을 조합하는 방법으로는 평균을 취하거나 (averaging) 최다득표자를 취하는 방법 (major voting)이 있다.

Bagging survival tree (Hothorn 등, 2004) 방법은 부스트랩표본 (boostrap sampling)을 반복적으로 생성하여 각각의 표본으로부터 생존나무를 구성하여, 새로운 개체에 대한 생존함수의 카플란-마이어 추정량을 이 새로운 개체가 각각의 부스트랩 표본에서 만들어진 생존나무의 마디에 속하는 모든 표본을 합한 표본으로부터 추정한다. 결과적으로 다수의 생존나무의 결과에서 도출된 마디들이 속하는 표본들은 새로운 개체와 유사한 개체들로 구성된 표본으로 이 모든 표본들을 합하여 추정된 생존함수의 추정량은 신뢰도가 높아지는 결과를 얻는다. Bagging survival tree를 수행하기 위해서는 (https://cran.rproject.org/web/packages/ipred/)를 참조하라.

Random survival forest 방법은 bagging과 마찬가지로 부스트랩 표본을 반복적으로 생성하여 각각의 표본으로부터 생존나무를 구성하여 평균을 취하는 점에서 동일하지만 bagging 과 다른 점은 생존나무를 구성할 때 전체 변수들 중 일부를 선택하여 사용한다는 것이다. 이러한 논리의 근거는 부스트랩 표본을 반복하게 되면 거의 67%의 표본이 원래의 표본과 겹치게 되어 bagging 과정에서 나온 결과들 간의 상관성이 매우 높아지게 된다. 따라서 bagging 방법에서 나타나는 결과들 간의 상관성을 줄이기 위하여 각 표본마다 다른 변수들을 선택하여 생존나무를 구성하며 이 결과들을 합쳐서 추론하는 것이다. 이 때 변수선택의 기준, 마디 수와 나무의 크기 등과 같은 것을 조절하는 과정에 대하여 Ishwaran 등 (2011)의 논문에 자세하게 연구되어 있다. Random survival forest를 수행하기 위해서는 (https://cran.rproject.org/web/packages/randomSurvivalForest/)를 참조하라.

Boosting은 앞에서 논의한 bagging과 random survivsl forest와 같이 다수의 결과를 반복적으로 산출하는데 bagging과 같이 독립적으로 표본을 생성하는 것이 아니라 축차적 (sequential)으로 전 단계에서 나온 추정과정을 이용하여 표본의 개체에 재가중치를 하면서 모형을 updating 하는 방법이다. Hothorn 등 (2006)의 논문에서는 중도절단된 생존자료에 대하여 로그선형모형에 대한 스코어 잔차에 gradient descent 알고리듬을 적용하여 축차적으로 표본을 재가중하면서 예측율을 높여가는 방법에 대하여 기술하고 있으며 boosting을 하지 않았을 경우와 random survival forests 방법과 비교하였다. 이미 앞 절에서 언급한 각스부스트도 이러한 방법을 활용한 것이다. 유전체 자료를 다루는 경우, 변수의 수가 많아질수록 앙상블의 방법이 예측오차의 분산을 줄이는 효과는 있지만 계산량이 급증하기 때문에 보다 효율적인 알고리듬을 개발하는 것도 중요한 이슈가 되고 있다.

6. 결론

지금까지 생존분석의 기초적인 개념과 확률이론에 대하여 소개하면서 전통적인 생존분석 방법부터 시작하여 유전체 자료를 분석하기 위하여 개발된 벌점함수를 이용한 통계적 방법론 및 기계학습과 접목되어 개발된 생존분석 방법들에 대하여 간단하게 살펴보았다. 생존분석은 중도절단된 자료를 포함하기 때문에 완전하게 관측된 자료를 기반으로 하는 기존의 통계적 방법들을 그대로 적용해서 분석할 수가 없다는 것이 가장 큰 특징이다. 생존함수를 추정하기 위하여 중도절단된 자료로부터 얻어지는 부분적인 정보를 반영하여 비모수적인 방법으로 카플란-마이어 추정량을 구하였고 서로 다른 두 그룹의 생존함수를 비교하기 위하여 순위의 개념에 기반하여 로그-순위 검정통계량을 제안하였다. 생존시간과 연관성이 있는 위험인자들을 찾아내기 위하여 칵스 회귀모형을 제안하였는데, 이 모형에서 위험인자들의 효과를 추정하거나 유의성을 검정하는 통계적인 추론 과정에서 중도절단된 자료들의 정보들이 반영되도록 통계적 방법들이 개발되었다. 특히 본 논문에서는 생존시간과 관련된 유전체 자료분석방법에 초점을 맞추어 벌점화를 이용하여 수많은 유전정보로부터 생존시간에 유의한 연관성이 있는 유전정보를 찾아내고 기계학습방법을 생존분석방법에 활용하여 예측력을 높이는 기법들에 대하여 개략적으로 살펴보았다. 개인의 유전체 자료가 점점 더 다양한 첨단생물학 기술로부터 동시에 얻어지고 비용이 급격하게 절감되면서 개인맞춤의학 (personalized medicine)에 유전체 자료분석을 위한 생존통계예측모형은 매우 중요한 기초기반 기술이 될 것이다. 따라서 본 논문에서 소개하였던 기초적인 개념과 방법들을 기반으로 생존분석에서 다루는 연구방법들을 이해하는데 큰 도움이 되었기를 바라고, 빅데이터로 수집되는 유전체 자료를 분석하기 위한 생존분석방법이 더 많이 개발되기를 기대하면서 본 논문을 맺고자 한다.

References

Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics*, **6**, 701-726.

Aalen, O. O. (1989). A linear regression model for the analysis of lifetimes. *Statistics in Medicine*, **8**, 907-925.

Aalen, O. O. (1993). Further results on the nonparametric linear regression model in survival analysis. *Statistics in Medicine*, **12**, 1569-1588.

Alizadeh, A., Eisen, M., Davis, R. E., Ma, C., Losso, I., Rosenwal, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., Marti, G., Moore, T., Hudsom, J., Lu, L., Lewis, D., Tibshirani, R., Sherlock, G., Chan, W., Greiner, T., Weisenburger, D., Armitage, K., Levy, R., Wilson, W., Greve, M., Byrd, J., Botstein, D., Brown, P. and Staudt, L. (2000). Identification of molecularly and clinically distinct subtypes of diffuse large B cell lymphoma by gene expression profiling. *Nature*, **403**, 503-511.

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1993). *Statistical models based on counting processes*, Springer-Verlag.

Ayer, T., Alagoz, O., Chhatwal, J., Shavlik, J. W., Kahn, C. E. and Burnside, E. S. (2010). Breast cancer risk estimation with artificial neural networks revisited: Discrimination and calibration. *Cancer*, **116**, 3310-3321.

Beer, D. G., Chen, G., Gharib, T. G., Giordano, T. J., Hayasaka, S., Huang, C. C., Iannettoni, M. D., Kardia, S. L. R., Kuick, R., Levin, A. M., Lin L., Lizyness, M. L., Misek, D. E., Orringer, M. B., Hanash, S., Taylor, J. M.G. and Thomas, D. G.(2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*, **8**, 816-824.

Benjamini, Y. and Hochberg, Y. (1995). Controlling th false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, **57**, 289-300.

Binder, H. and Schumacher, M. (2008). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics*, **9**, 1471-2105.

Breslow, N. E. (1975). Analysis of survival data under the proportional hazards model. *International Statistical Review*, **43**, 45-57.

Breslow, N. E. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censoring. *Annals of Statistics*, **2**, 437-453.

Brieman, L. (1998). Arcing classifiers (with discussion). *Annals of Statistics*, **26**, 801-849.

Brieman, L., Friedman, J., Stone, C. J. and Olshen, R. A. (1984). *Classification and regression trees*, Taylor & Francis.

Buhlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, **22**, 477-505.

Chen, Z. and Zhang, G. (2016). Comparing survival curves based on medians. *BMC Medical Research Methodology*, **16**, 1-7.

Ciampi, A., Thiffault, J. Nakache, J. and Asselain, B. (1986). Stratification by stepwise regression, correspondence analysis and recursive partition: A comparison of three methods of analysis for survival data with covariates. *Computational Statistics and Data Analysis*, **4**, 185-204.

Cox, D. R. (1972). Regression models and life tables (with Discussion). *Journal of the Royal Statistical Society B*, **34**, 187-220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**, 269-276.

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals (with Discussion). *Journal of the Royal Statistical Society B*, **30**, 248-275.

Cruz, J. A. and Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics*, **2**, 59-77.

Dudoit, S., Yang, Y., Callow, M. and Speed, T. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111-139.

Efron, B. (1967). The two sample problem with censored data. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, New York: Prentice-Hall*, **4**, 831-853.

Eisen, M., Spellman, P., Brown, P. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Nat. Acad. Sci.*, **95**, 14863-14868.

Fisher, L. and Kanarek, P. (1974). Presenting censored survival data when censoring and survival times may not be independent. *Reliability and Biometry: Statistical Analysis of Lifelength (Proschan F. and Serfling, R. J. Eds.), Philadelphia: SIAM*, 303-326.

Fleming, T. R. and Harrington, D. P. (1991). *Counting processes and survival analysis*, John Wiley and Sons, New York.

Freund, Y. and Schapire, R. (1995). A decision - Theretical generalization of on-line learning and an application to boosting. *Proceedings of the Second European Conference on Computational Learning Theory*, Springer, Berlin.

Friedman, J., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion). *J Annals of Statistics*, **28**, 337-407.

Garber, M. E., Troyanskaya, O. G., Schluens, K., Petersen, S., Thaesler, Z., PacynaGengelbach, M., van de Rjin, M., Rosen, G. D., Perou, C. M., Whyte, R. I., Altman, R. B., Brown, P. O., Botstein, D. and Petersen, I. (2001). Diversity of gene expresssion in adenocarcinoma of the lung. *Proceesings of National Academic Sciences*, **98**, 13784-13789.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M., A., Bloomfield, C. D. and Lander, E. S. Hastie, T. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-537.

Greenwood, M. (1926). The natural duration of cancer. *In Reports On Public Health and Medical Subjects, London: His Majesty's Stationery Office*, **33**, 1-26.

Hastie, T., Tibshirani, R. and Wainwright, M. (2016). *Statistical learning with sparsity: The lasso and generalizations*, CRC Press.

Hoerl, A and Kennard, R. (1988). Ridge regression. *In Encyclopedia of Statistical Sciences* , **8**, 129-136.

Hothorn, T., Buhlmann, P., Dudoit, S., Molinaro, A. and van der Laan, M. J. (2006). Survival ensemble. *Biostatistics*, **7**, 355-373.

Hothorn, T., Lausen, B., Benner, A. and Radespiel-Troger, M. (2004). Bagging survival trees. *Statistics in Medicine* , **23**, 77-91.

Ishwaran, H., Kogalur, U. B., Chen, X. and Andy, J.M. (2011). Random survival forests for high-dimensional data. *Statistical Analysis and Data Mining*, **4**, 115-132.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457-481.

Khan, F. M. and Zubek, V. B. (2008). Support vector regression for censored data (SVRc): A novel tool for survival analysis. *Eighth IEEE International Conference on Data Mining*.

Kim, J. H. (2016). *Introduction to survival analysis with R*, Freedom Academy.

Klein, J. P. and Moeschberger, M. L. (2010). *Survival analysis: Techniques for censored and truncated data*, 2nd Ed, Springer.

Korn, J. M., Kuruvilla, F. G. and McCarroll, S. A., *et al.* (2008). Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature Genetics*, **40**, 1253-1260.

Lee, T. S. and Lee, M. J. (2017). Analysis of stage III proximal colon cancer using the Cox proportional hazards model. *Journal of the Korean Data & Information Science Society* , **28**, 349-359.

Lee, M., Kuo, F. C., Whitmore, G. A. and Sklar, J. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *PNAS*, **97**, 9834-9839.

Leblanc, M. and Crowley, J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association*, **88**, 457-467.

Lin, D.Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, **81**, 61-71.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, **14**, 945-965.

Newton, M., Kendziorski, C., Richmond, C., Blatter, F. and Tsui, K. (2000). On differential variability of expression ratios: Imporving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, **8(11)**, 37-52.

Sebat, J., Lakshmi, B. and Malhotra, D. *et al.* (2007). Strong association of de novo copy number mutations with autism. *Science*, **316**, 445-449.

Scharfstein, D. O. and Robins, J. M. (2002). Estimation of the failure time distribution in the presence of informative censoring. *Biometrika*, **89**, 617-634.

Shivaswamy, P. K., Chu, W. and Jansche, M. (2007). A support vector approach to censored targets. *Seventh IEEE International Conference on Data Mining*, **93**, 655-660.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate desent. *Journal of Statistical Software* **39** 1-13.

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E. and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate ii: Radical prostatectomy treated patients. *Journal of Urology*, **16**, 1076-1083.

Theodoridis, S. and Konstantinos, K. (2008). *Pattern Recognition*, 4th Ed.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine* , **16**, 385-395.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society B*, **67**, 91-108.

Tusher, G. V., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, **98**, 5116-5121.

Wahba, G. (1980). Spline bases, regularization, and generalized cross-validation for solving approximation problems with large quantities of niosy data. *Proceesings of the International Conference on Approximation theory in honour of George Lorenz, Academic Press, Austin, Texas*.

Yoo, H. (2018). Prediction model for clustered survival data with missing covariates using decision tree. *Journal of the Korean Data & Information Science Society*, **29**, 1119-1126.

Zhang, H. H. and Lu, W. (2007). Adaptive lasso for Cox's proportional hazards model. *Biometrika*, **94**, 691-703.

Zou, H. and Hastie T. (2005). Regularization and variable selection via the elastic net. *Journal of the*

Royal Statistical Society B, **67** 301-320.

Journal of the Korean Data & Information Science Society
2018, **29**(6), 1391–1408

http://dx.doi.org/10.7465/jkdi.2018.29.6.1391
한국데이터정보과학회지

Review of the survival analysis methods for genetic data

Seungyeoun Lee[1]

[1]Department of Mathematics and Statistics, Sejong University

Received 15 October 2018, revised 15 November 2018, accepted 16 November 2018

Abstract

Survival analysis focuses on the statistical inference for the time to event of interest, which cannot be often completely observed due to censoring. Considering the characteristics of these censored data, traditional survival analysis methods have been developed for estimation, testing, and model development to predict survival time for patients based on clinical data. However, large-scale data from high-throughput genomic technologies, especially microarrays, have been collected, which poses the challenging statistical issues in combining those with the survival time. Many statistical methods have been developed by additionally considering the high-dimensional genomic information in the statistical prediction model constructed only by the existing clinical data. Recently, there have been many studies on the methodology of integrating different types of genomic data through various advanced biologic techniques, which results in making an early prediction for the disease and developing personalized medicine. As well, there has been considerable interest in applying machine learning techniques to analyse these complex and huge amount of genomic data associated with the censored data. In this paper, we review the basic concepts in survival analysis, traditional statistical methods based on clinical data, more appropriate statistical methods dealing with genomic data, and machine learning methods extended to the survival analysis.

Keywords: Censoring, machine learning, nonparametric methods, penalty function, statistical predictive model, survival time.

[1] Corresponding author: Professor, Department of Mathematics and Statistics, Sejong University, Seoul, 05006, Korea. E-mail : leesy@sejong.ac.kr