Journal of the Korean Data & Information Science Society
2012, **23**(4), 739–748

http://dx.doi.org/10.7465/jkdi.2012.23.4.739
한국데이터정보과학회지

생존자료분석을 위한 혼합효과 최소제곱 서포트벡터기계[†]

황창하[1] · 심주용[2]

[1]단국대학교 정보통계학과 · [2]인제대학교 데이터정보학과

접수 2012년 6월 14일, 수정 2012년 7월 3일, 게재확정 2012년 7월 9일

요약

최소제곱 서포트벡터기계 (least squares support vector machine)는 분류 및 비선형 회귀분석에서 유용하게 사용되고 있는 통계적 기법이다. 본 논문에서는 각 집단별로 생존자료가 관측된 경우 적용할 수 있는 LS-SVM을 제안한다. 제안된 모형은 임의우측 중도절단자료를 비선형 회귀모형에 적용할 수 있게 Kaplan- Meier의 중도절단분포의 추정값을 이용하여 구해진 가중값을 사용하고, 집단 간의 변동을 나타내기 위하여 임의효과항을 포함한다. 벌칙상수와 커널모수의 최적값을 구하기 위하여 일반화 교차타당성함수가 사용되고 모의실험에서는 임의효과항을 포함하지 않은 LS-SVM과 성능을 비교함으로써 제안된 방법의 우수성을 보이기로 한다.

주요용어: 벌칙상수, 일반화 교차타당성함수, 임의우측 중도절단자료, 임의효과, 중도절단분포, 최소제곱 서포트벡터기계, 커널모수, 혼합효과모형.

1. 서론

일반적으로 생존분석은 양의 값 (positive value)을 갖는 확률변수를 분석하는 통계적 방법으로 알려져 있다. 즉 생존분석이란 같은 상태를 유지하고 있는 시간의 길이를 분석하고, 이 시간의 길이에 영향을 미치는 원인을 분석하는 통계적 방법이다. 전통적으로 통계학이나 의학 분야에서 주로 사용되어 왔으나, 최근 생존분석이라는 용어 대신 사건까지의 시간분석 (time to event analysis; Moulton과 Dibley, 1997) 등의 포괄적인 용어가 사용되어, 경제학에서는 취업기간이나 실업기간 또는 회사의 도산시간의 분석에, 심리학에서는 학습인지시간 분석에, 공학에서는 제품의 평균수명 추정에 응용되고 있다. 생존자료는 특성상 중도절단자료 (censored data)를 주로 포함하므로 일반적인 통계적 방법과는 다른 자료 분석의 방법이 제안되어졌다. 중도절단의 양상은 크게 세 가지 형태로 나눌 수 있는데, 본 연구에서는 종료시각의 임의 설정에 의해 발생되는 우측 중도절단 (right censoring)의 형태만을 포함한다. 생존분석은 생명표를 제외하면 Kaplan과 Meier의 비모수 분석법, Cox의 비례위험 (proportional hazard) 회귀모형 그리고 가속화 고장시간모형 (accelerated failure time model; Buckley와 James, 1979)의 개발 이후 많은 발전을 보이고 있으며, 생존분석에 대한 기본적인 내용은 참고문헌 Cox (1972), Kaplan과 Meier (1958), Kalbfleisch와 Prentice (1980), Miller (1981)에 설명되어 있다. 본 연구에서 우리는 생존분석에서 입력벡터 (공변량)의 생존시간에 대한 영향력을 분석하는 회귀모형인 가속화 고장시간모형에 기반을 둔 새로운 모형을 제안하고자 한다.

[†] 이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임 (2011-0027344).
[1] (448-701) 경기도 용인시 수지구 죽전동 126번지, 단국대학교 정보통계학과, 교수.
[2] 교신저자: (621-749) 경남 김해시 어방동, 인제대학교 데이터정보학과, 겸임교수.
E-mail: ds1631@hanmail.net

혼합효과모형은 고정효과 (fixed effect)와 임의효과를 모두를 이용하여 종속변수와 입력변수들의 관계를 설명하는 모형이다. 고정효과란 종속변수를 관측할 때 부가적으로 관측되는 입력변수들과 관련된 효과를 의미하고 임의효과란 각 집단 내에서는 고정적이지만 주어진 실험에서는 관측되지 않은 어떤 요인들에 의해 임의적으로 (randomly) 발생되는 집단 간의 변동효과를 의미한다. Harville (1976, 1977)과 Laird와 Ware (1982)는 선형혼합모형 (linear mixed effects model)을 제안하였고, McCulloch와 Searle (2000)은 선형혼합모형과 일반화선형모형 (GLM)을 결합한 일반화선형혼합모형 (generalized linear mixed model)을 제안하였다. 혼합모형에서 종속변수와 입력변수들의 관계를 더 잘 설명하기위해 비선형혼합모형 (nonlinear mixed effects model)이 많이 이용되는데, 주로 선형혼합모형의 일반화된 형태인 2단계계층형태 (two stage hierarchical form)로 주어진다 (Davidian과 Giltinan, 1995; Vonesh와 Chinchilli, 1996). 또한 비모수적인 방법으로서 평활스프라인회귀 (smoothing spline regression)가 비선형혼합모형의 분석에 많이 사용되고 있다 (Green과 Silverman, 1994; Wahba, 1990).

서포트벡터기계 (support vector machine; SVM)는 1995년에 Vapnik 중심의 MIT 인공지능 연구실에서 개발되어 통계학, 생물정보학, 금융정보학, 데이터마이닝, 컴퓨터과학 및 정보과학 등의 분야에서 분류 및 회귀함수추정을 위한 우수한 성능의 기법으로 많이 활용되고 있다. 그리고 SVM의 기본원리인 커널기법을 사용하여 선형 모형을 비선형 모형으로 자연스럽게 확장하고 우수한 분류 및 회귀함수추정의 성능을 보여주는 커널기계 (kernel machine)도 SVM과 함께 분류 및 회귀문제를 위해 많이 활용되고 있다. SVM은 원래 집단이 두 개인 경우 통계학적 이론을 배경으로 오분류율 (misclassification rate)을 최소화 시키는 최적분리 초평면 (hyperplane)을 제공함으로써 다양한 응용분야의 분류 및 회귀문제에서 우수한 성능이 입증되고 있다. SVM은 커널함수 (kernel function)를 사용하는 비선형 모형으로서 기존의 선형 모형을 비선형 모형으로 쉽게 확장할 수 있는 이론적 근거를 제공하고 있다. SVM 및 커널기계가 주목받는 이유는 첫째, 명백한 이론적 근거에 기반을 두므로 결과 해석이 용이하고, 둘째, 실제 응용에 있어서 인공신경망 보다 더 좋은 성과를 내고, 셋째, 적은 학습 자료만으로 신속하게 분류 및 회귀학습을 수행할 수 있기 때문이다. SVM 및 커널기계에 대한 자세한 내용은 참고문헌 Cristianini와 Shawe-Taylor (2000), Gunn (1998), Scholkopf와 Smola (2002), Vapnik (1995, 1998)에 설명되어 있다. Suykens와 Vandewalle (1999)은 능형회귀 (ridge regression; Sanuders 등, 1998) 개념을 도입하여 SVM의 부등식 제한조건을 등식 제한조건으로 변환함으로써 최소제곱 서포트벡터기계 (least squares-SVM; LS-SVM)을 제안하였다. LS-SVM은 수행 능력이 SVM에 비슷할 뿐만 아니라 이차프로그래밍 문제를 선형방정식 (linear equations)문제로 해결하여 훈련시간을 상당히 줄일 수 있는 장점이 있다. SVM, 커널기계 및 LS-SVM 의 응용에 대한 자세한 내용은 참고문헌 Kim 등 (2008), Shim과 Lee (2009), Jo 등 (2010), Hwang과 Shim (2011)에 설명되어 있다.

각 집단 (group)내의 개체들의 생존시간을 측정하여 얻어지는 경우, 우리는 중도절단된 자료를 포함하는 복잡한 구조를 갖는 생존자료 및 집단화된 생존자료를 분석하기 위한 LS-SVM 기반 통계모형 및 기법을 개발하고 그 응용방법을 도출하는데 있다. 우리는 집단 간의 변동효과를 추정할 수 있도록 임의효과 (랜덤효과) 항을 도입하여 LS-SVM에 혼합효과모형을 결합한 모형을 제안한다. 생존분석을 위한 LS-SVM 기반 모형의 장점은 크게 세 가지이다. 첫째, 간단한 볼록최적화 (convex optimization) 기법을 사용하여 비교적 쉽게 추정값을 구할 수 있다. 둘째, 커널함수를 사용하기 때문에 예측력이 뛰어난 비선형 모형일 뿐만 아니라 원소간 곱 (componentwise product) 커널을 사용하면 비선형 모형이지만 생존분석에서 특히 중요하게 요구되는 해석 가능한 모형이 된다. 셋째, 특히 고차원 (high dimensional) 자료에 대해서 성능이 우수한 생존분석 모형이다.

본 논문은 다음과 같이 구성되어 있다. 2절에서는 생존자료분석을 위한 LS-SVM에 대해서 3절에서는 생존자료분석을 위한 임의효과항을 포함하는 혼합효과 LS-SVM에 대해서 기술한다. 4절에서는

각각 모의실험에 대한 결과를 설명하고 5절은 결론을 기술한다.

2. 최소제곱 서포트벡터기계

주어진 입력벡터를 $\boldsymbol{x}_i$라고 할 때, 우리는 다음과 같은 회귀모형을 가정한다.

$$m(\boldsymbol{x}_i) = b_0 + \boldsymbol{w}'\boldsymbol{\phi}(\boldsymbol{x}_i),\ i = 1, \cdots, n, \tag{2.1}$$

여기서 b_0는 편의항, $\boldsymbol{\phi}(\boldsymbol{x}_i)$는 $d_f \times 1$ 비선형특징사상함수 (nonlinear feature mapping function), 그리고 $\boldsymbol{w}$는 $\boldsymbol{\phi}$에 대응되는 모수벡터이다. $\boldsymbol{\phi}$에 대해 $\boldsymbol{\phi}(\boldsymbol{x})'\boldsymbol{\phi}(\boldsymbol{z}) = K(\boldsymbol{x}, \boldsymbol{z})$를 만족하는 커널함수가 존재한다 (Mercer, 1909).

실제로, 우리들은 종속변수인 생존시간 t_i를 관측할 수는 없고 관측변수 $y_i = \min(t_i, c_i)$와 $\delta_i = I(t_i \leq c_i)$를 관측할 수 있다. 여기서, c_i는 $\boldsymbol{x}_i$에 대응하는 중도절단변수 (censoring variable)이다. 중도절단변수들은 임의우측 중도절단 (randomly right censored) 되었고 서로 독립으로 가정된다. 대부분의 경우에 중도절단변수들의 생존함수는 알려지지 않으므로 주로 Kaplan-Meier (1958) 추정량 또는 그 변종 (variation)에 의해 추정된다. 우리는 다음과 같은 생존함수 G의 추정값을 사용하고자 한다.

$$\widehat{G}_{(y)} = \begin{cases} \displaystyle\prod_{i:y_{(i)} \leq y} \left(\frac{n-i}{n-i+1} \right)^{1-\delta_{(i)}} & \text{만약 } y \leq y_{(n)} \\ 0 & \text{그 외} \end{cases} \tag{2.2}$$

여기서 $(y_{(i)}, \delta_{(i)})$는 (y_i, δ_i)의 y_i $(i = 1, \cdots, n)$를 기준으로 정렬된 형태이다.

우리는 Koul 등 (1981)이 제안한 가중값 u_i를 이용하여 다음과 같은 LS-SVM 형태의 최적화 문제를 고려한다.

$$min \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + \frac{C}{2}\sum_{i=1}^{n} u_i e_i^2 \ over\ \{\boldsymbol{w}, b_0, \boldsymbol{e}\} \tag{2.3}$$

제약조건:

$$y_i - (b_0 + \boldsymbol{w}'\boldsymbol{\phi}(\boldsymbol{x}_i)) = e_i,\ i = 1, \cdots, n, \tag{2.4}$$

여기서 $u_i = \delta_i / G(y_i)$, $y_i = \min(t_i, c_i)$ 그리고 $\delta_i = I(t_i \leq c_i)$이다. 우리는 식 (2.3)과 (2.4)를 이용하여 라그랑즈함수 (Lagrange function)를 다음과 같이 만들 수 있다.

$$L(\boldsymbol{w}, b_0, e_i : \alpha_i) = \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + \frac{C}{2}\sum_{i=1}^{n} u_i e_i^2 - \sum_{i=1}^{n} \alpha_i (e_i - (y_i + b_0 - \boldsymbol{w}'\boldsymbol{\phi}(\boldsymbol{x}_i))) \tag{2.5}$$

여기서 α_i는 라그랑즈배수이다. 식 (2.5)를 $(\boldsymbol{w}, b_0, e_i)$에 관하여 편미분하면 다음과 같은 결과를 얻는다.

$$\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i \boldsymbol{\phi}(\boldsymbol{x_i}),\ \sum_{i=1}^{n} \alpha_i = 0,\ \alpha_i = C u_i e_i,\ i = 1, \cdots, n. \tag{2.6}$$

식 (2.6)에서 우리는 $u_i = 0$ 인 경우 $\alpha_i = 0$가 성립함을 알 수 있다. 다시 식 (2.5)를 α_i에 관하여 편미분하면 다음과 같은 결과를 얻는다.

$$y_i - b_0 - K_i\boldsymbol{\alpha} - \alpha_i/(Cu_i) = 0,\ i \in \mathrm{I}_s = \{i = 1, \cdots, n | u_i \neq 0\} \tag{2.7}$$

여기서 $K_i = K(\boldsymbol{x}_i, \boldsymbol{x})$는 $\boldsymbol{x}_i$에 대응하는 커널함수행렬 K의 행벡터이고 커널함수행렬 K는 $\{\boldsymbol{x}_i, i = 1, \cdots, n\}$로부터 구해진 $n \times n$ 행렬, 그리고 $\boldsymbol{\alpha}' = (\alpha_1, \alpha_2, \cdots, \alpha_n)'$는 $n \times 1$ 라그랑즈배수 벡터이다.

그러므로 b_0와 α_i ($i \in \mathrm{I_s}$)의 추정값은 다음의 선형방정식에서 구해진다.

$$\begin{pmatrix} 0' & 1'_{n_s} \\ \mathbf{1}_{n_s} & K_s + \dfrac{1}{C} diag(\boldsymbol{u}_s)^{-1} \end{pmatrix} \begin{pmatrix} b_0 \\ \boldsymbol{\alpha}_s \end{pmatrix} = \begin{pmatrix} 0 \\ \boldsymbol{y}_s \end{pmatrix} \tag{2.8}$$

여기서 n_s는 I_s의 길이, $\boldsymbol{\alpha_s} = \{\alpha_i, i \in \mathrm{I_s}, i = 1, \cdots, n\}$, $\boldsymbol{y}_s = \{y_i, i \in \mathrm{I_s}, i = 1, \cdots, n\}$, $\boldsymbol{u}_s = \{u_i, j \in \mathrm{I_s}, i = 1, \cdots, n\}$, 그리고 $K_s = K(\boldsymbol{x_s}, \boldsymbol{x_s})$이다. 따라서, 입력벡터 $\boldsymbol{x}_i$에 대한 회귀함수의 추정값은 다음과 같이 구해진다.

$$m(\boldsymbol{x}_i) = b_0 + K(\boldsymbol{x}_i, \boldsymbol{x}_s)\boldsymbol{\alpha}_s. \tag{2.9}$$

이제 벌칙상수와 커널모수를 결정하는 모형선택 (model selection) 문제를 생각한다. 생존분석을 위한 LS-SVM의 성능은 벌칙상수와 커널모수의 값에 영향을 받으므로 최적의 벌칙상수와 커널모수의 값을 선택하여야 한다. 먼저 다음과 같은 LOO (leave-one-out) 교차타당성 (cross validation) 함수를 고려한다.

$$CV(\theta) = \frac{1}{n_s} \sum_{i=1}^{n} u_i (y_i - m_\theta{}^{(-i)}(\boldsymbol{x}_i))^2. \tag{2.10}$$

여기서 θ는 벌칙상수와 커널모수로 이루어진 벡터이고, $m_\theta^{(-i)}(\boldsymbol{x}_i)$는 전체자료 중 i번째 자료를 제외한 나머지 자료들을 이용하여 구한 $m_\theta(\boldsymbol{x}_i)$의 추정값이다. 위의 교차타당성함수는 0이 아닌 u_i와 관련된 전체자료 중 k번째 자료를 $\{y_k, \boldsymbol{x}_k, u_k\}$로 표기함으로써 다음과 같이 표현될 수도 있다.

$$CV(\theta) = \frac{1}{n_s} \sum_{k=1}^{n_s} u_k (y_k - m_\theta^{(-k)}(\boldsymbol{x}_k))^2, \tag{2.11}$$

여기서 $m_\theta^{(-k)}(\boldsymbol{x}_k)$는 0이 아닌 u_i와 관련된 전체자료 중 k번째 자료를 제외한 나머지 자료들을 이용하여 구한 $m_\theta(\boldsymbol{x}_k)$의 추정값이다. 주어진 θ의 값에 대하여 n_s 개의 $m_\theta^{(-k)}(\boldsymbol{x}_k)$가 구해져야 하므로 식 (2.11)의 LOO 교차타당성함수를 이용하여 $m_\theta^{(-k)}(\boldsymbol{x}_k)$의 적합한 값을 구하는 것은 계산적으로 매우 비효율적이다.

입력벡터 $\boldsymbol{x}_k$에 대응되는 회귀함수의 추정값이 $m(\boldsymbol{x}_k) = S(\boldsymbol{x}_k, \boldsymbol{x}_s)\boldsymbol{y}_s$로 표현될 수 있으므로, LOO lemma (Craven과 Wahba, 1979)를 이용하여 일반화교차타당성 (generalized cross validation)함수는 다음과 같이 구해진다.

$$GCV(\theta) = \frac{1}{n_s} \frac{\sum_{k=1}^{n_s} u_k (y_k - m_\theta(\boldsymbol{x}_k))^2}{(1 - tr(S)/n_s)^2} \tag{2.12}$$

여기서 $S = (\mathbf{1}_{n_s} S_{12} + K_s S_{22})$이고, $1 \times n_s$ 행렬 S_{12}와 $n_s \times n_s$ 행렬 S_{22}는 식 (2.8)의 왼쪽 행렬의 역행렬의 부분행렬 (submatrix)들이다.

$$\begin{pmatrix} 0' & 1'_{n_s} \\ \mathbf{1}_{n_s} & K_s + \dfrac{1}{C^2} diag(\boldsymbol{u}_s)^{-1} \end{pmatrix}^{-1} = \begin{pmatrix} S_{11} & S_{12} \\ S'_{12} & S_{22} \end{pmatrix}$$

3. 혼합효과 최소제곱 서포트벡터기계

각 집단내의 개체들의 생존시간을 측정하여 얻어지는 자료의 경우 입력벡터 $\boldsymbol{x}_{ij}$에 대해 회귀함수 $m(\boldsymbol{x}_{ij})$가 다음과 같은 형태로 관련되어 있다고 가정한다.

$$m(\boldsymbol{x}_{ij}) = b_0 + \boldsymbol{w}'\boldsymbol{\phi}(\boldsymbol{x}_{ij}) + b_i,\ j = 1, \cdots, n_i,\ i = 1, \cdots, n, \tag{3.1}$$

여기서 n 은 집단의 수, n_i는 i번째 집단의 자료수, b_i는 i번째 집단과 다른 집단과의 변동 (variation)을 나타내는 임의효과 모수, $\boldsymbol{\phi}(\boldsymbol{x}_{ij})$는 $d_f \times 1$ 비선형특징사상함수, 그리고 $\boldsymbol{w}$는 $\boldsymbol{\phi}(\boldsymbol{x}_{ij})$에 대응되는 모수벡터이다. 이 절에서는 먼저 식 (2,3)과 (2.4)를 이용하여 다음과 같은 LS-SVM 형태의 최적화 문제를 고려한다.

$$min \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + \frac{C}{2}\sum_{i=1}^{n}\sum_{j=1}^{n_i} u_{ij}e_{ij}^2 \text{ over } \{\boldsymbol{w}, b_0, \boldsymbol{b}, \boldsymbol{e}\} \tag{3.2}$$

제약조건:

$$y_{ij} - (b_0 + \boldsymbol{w}'\boldsymbol{\phi}(\boldsymbol{x}_{ij})) = e_{ij},\ j = 1, \cdots, n_i,\ i = 1, \cdots, n \tag{3.3}$$

여기서 $u_{ij} = \delta_{ij}/G_i(y_{ij})$, $y_{ij} = \min(t_{ij}, c_{ij})$ 그리고 $\delta_{ij} = I(t_{ij} \le c_{ij})$이다. 우리는 식 (3.2)와 (3.3)을 이용하여 라그랑즈함수를 다음과 같이 만들 수 있다.

$$L(\boldsymbol{w}, b_0, e_{ij} : \alpha_{ij}) = \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + \frac{C}{2}\sum_{i=1}^{n}\sum_{j=1}^{n_i} u_{ij}e_{ij}^2 - \sum_{i=1}^{n}\sum_{j=1}^{n_i}\alpha_{ij}(e_{ij} - (y_{ij} + b_0 - \boldsymbol{w}'\boldsymbol{\phi}(\boldsymbol{x}_{ij}))) \tag{3.4}$$

여기서 α_{ij}는 라그랑즈배수이다. 식 (3.4)를 $(\boldsymbol{w}, b_0, e_{ij})$에 관하여 편미분하면 다음과 같은 결과를 얻는다.

$$\boldsymbol{w} = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\alpha_{ij}\boldsymbol{\phi}(\boldsymbol{x}_{ij}),\ \sum_{i=1}^{n}\sum_{j=1}^{n_i}\alpha_{ij} = 0,\ \alpha_{ij} = Cu_{ij}e_{ij},\ j = 1, \cdots, n_i,\ i = 1, \cdots, n. \tag{3.5}$$

식 (3.5)에서 우리는 $u_{ij} = 0$ 인 경우 $\alpha_{ij} = 0$가 성립함을 알 수 있다. 다시 식 (3.4)를 α_{ij}에 관하여 편미분하면 다음과 같은 결과를 얻는다.

$$y_{ij} - b_0 - K^{ij}\boldsymbol{\alpha} - \alpha_{ij}/(Cu_{ij}) = 0,\ j \in \mathrm{I_s(i)} = \{j = 1, \cdots, n_i | u_{ij} \neq 0\} \tag{3.6}$$

여기서, $N = \sum_{i=1}^{n} n_i$, K^{ij}는 $\boldsymbol{x}_{ij}$에 대응하는 커널함수행렬 K의 행벡터이고 커널함수행렬 K는 $\{\boldsymbol{x}_{ij}, j = 1, \cdots, n_i, i = 1, \cdots, n\}$로부터 구해진 $N \times N$ 행렬, 그리고 $\boldsymbol{\alpha} = (\alpha_{11}, \alpha_{12}, \cdots, \alpha_{n,n_n})'$는 $N \times 1$ 라그랑즈배수 벡터이다. 그러므로 b_0와 α_{ij} $(j \in \mathrm{I_s(i)})$의 추정값은 다음의 선형방정식에서 구해진다.

$$\begin{pmatrix} 0' & 1'_{N_s} \\ \mathbf{1}_{N_s} & K_s + \dfrac{1}{C}diag(\boldsymbol{u}_s)^{-1} \end{pmatrix} \begin{pmatrix} b_0 \\ \boldsymbol{\alpha}_s \end{pmatrix} = \begin{pmatrix} 0 \\ \boldsymbol{y}_s \end{pmatrix} \tag{3.7}$$

여기서 $n_{s(i)}$는 $I_{s(i)}$의 길이, $N_s = \sum_{i=1}^{n} n_{s(i)}$, $\boldsymbol{\alpha_s} = \{\alpha_{ij}, j \in \mathrm{I_s(i)}, i = 1, \cdots, n\}$, $\boldsymbol{y_s} = \{y_{ij}, j \in \mathrm{I_s(i)}, i = 1, \cdots, n\}$, $\boldsymbol{u_s} = \{u_{ij}, j \in \mathrm{I_s(i)}, i = 1, \cdots, n\}$, 그리고 $K_s = K(\boldsymbol{x}_s, \boldsymbol{x}_s)$이다.

따라서, 속하는 집단을 고려하지 않을 때, 입력벡터가 $\boldsymbol{x}_{ij}$에 대응되는 회귀함수의 추정값은 다음과 같이 구해진다.

$$m(\boldsymbol{x}_{ij}) = b_0 + K(\boldsymbol{x}_{ij},\ \boldsymbol{x}_s)\boldsymbol{\alpha}_s. \tag{3.8}$$

식 (3.7)에서 구한 회귀함수의 추정값은 임의효과를 고려하지 않고 구해졌으므로 우리는 $r_{ij} = y_{ij} - m(\boldsymbol{x}_{ij})$ for $j \in I_{s(i)}$를 이용하여 임의효과항의 추정값을 다음과 같이 구한다.

$$b_i = \frac{1}{n_{s(i)}} \sum_{j \in I_{s(i)}} r_{ij}. \tag{3.9}$$

따라서 i번째 집단에 속하는 입력벡터 $\boldsymbol{x}_{ij}$에 대응되는 회귀함수의 추정값은 다음과 같이 구해진다.

$$m(\boldsymbol{x}_{ij}) = b_0 + b_i + K(\boldsymbol{x}_{ij}, \boldsymbol{x}_s)\boldsymbol{\alpha}_s. \tag{3.10}$$

모형선택을 위하여 먼저 다음과 같은 LOO 교차타당성함수를 고려한다.

$$CV(\theta) = \frac{1}{N_s} \sum_{i=1}^{n} \sum_{j \in I_{s(i)}} u_{ij}(y_{ij} - m_\theta^{(-ij)}(\boldsymbol{x_{ij}}))^2. \tag{3.11}$$

집단을 고려하지 않고 자료들을 표현할 때, 0이 아닌 u_{ij}와 관련된 전체자료 중 k번째 자료를 $\{y_k, \boldsymbol{x}_k, u_k\}$로 표기함으로써 위의 교차타당성함수는 다음과 같이 표현될 수도 있다.

$$CV(\theta) = \frac{1}{N_s} \sum_{k=1}^{N_s} u_k(y_k - m_\theta^{(-k)}(\boldsymbol{x}_k))^2, \tag{3.12}$$

여기서 $m_\theta^{(-k)}(\boldsymbol{x}_k)$는 0이 아닌 u_i와 관련된 전체자료 중 k번째 자료를 제외한 나머지 자료들을 이용하여 구한 $m_\theta(\boldsymbol{x}_k)$의 추정값이다. 만약 $\boldsymbol{x}_k$가 i번째 집단에서 관측되었다면 대응되는 회귀함수의 추정값은 식 (3.8)을 이용하여 다음과 같이 표현될 수 있다.

$$m(\boldsymbol{x}_k) = D_k(\boldsymbol{y}_s - S_k\boldsymbol{y}_s) + S_k\boldsymbol{y}_s = (D_k - D_kS_k + S_k)\boldsymbol{y}_s, \tag{3.13}$$

여기서 $D_k = \big(1/n_{s(i)}, \cdots, 1/n_{s(i)}, 0, \cdots, 0\big)$, $S_k = (S_{12}, K(\boldsymbol{x}_k, \boldsymbol{x}_s)S_{22})$이고, $1 \times N_s$ 행렬 S_{12}와 $N_s \times N_s$ 행렬 S_{22}는 식 (3.6)의 왼쪽 행렬의 역행렬의 다음과 같은 부분행렬들이다.

$$\begin{pmatrix} 0' & 1'_{N_s} \\ \mathbf{1}_{N_s} & K_s + \dfrac{1}{C} diag(\boldsymbol{u}_s)^{-1} \end{pmatrix}^{-1} = \begin{pmatrix} S_{11} & S_{12} \\ S'_{12} & S_{22} \end{pmatrix}.$$

한편 $m(\boldsymbol{x_k}) = (D_k - D_kS_k + S_k)\boldsymbol{y_s} = H_k\boldsymbol{y_s}$로 표현될 수 있으므로, LOO lemma (Craven과 Wahba, 1979)를 이용하면 일반화교차타당성함수는 다음과 같이 구해진다.

$$GCV(\boldsymbol{\theta}) = \frac{1}{N_s} \frac{\sum_{k=1}^{N_s} u_k \left(y_k - m_{\boldsymbol{\theta}}(\boldsymbol{x_k})\right)^2}{(1 - tr(H)/N_s)^2}. \tag{3.14}$$

4. 실험 및 결과

제안된 혼합효과 LS-SVM의 유한 표본에 대한 성능을 설명하기위해 모의실험을 수행한다. 혼합효과 LS-SVM과 임의효과항을 포함하지 않는 일반적 LS-SVM을 집단별로 생성된 모의 생존자료의 회귀함수 (각 입력 변수에 대응하는 생존시간의 평균)의 추정 성능을 비교함으로써 제안된 방법의 우수성을 보이기로 한다.

우리는 집단의 수를 6개로 정하고, 각 집단의 입력변수 x를 균일분포 U(0,6)에서 생성하였고, 생존시간, 중도절단시간 및 관측시간을 다음과 같이 생성하였다.

$$t_{ij} \sim N(f(x_{ij}), 0.5^2),\ c_{ij} \sim N(c_0 + f(x_{ij}), 0.5^2),\ f(x_{ij}) = 2 + b_i + 2sin(2\pi x_{ij}),$$
$$y_{ij} = min(t_{ij}, c_{ij}), \delta_{ij} = I(t_{ij} \le c_{ij}),\ i = 1, \cdots, 6,\ j = 1, \cdots, 20.$$

여기서 b_i는 평균이 0이고 분산이 4인 정규분포에서 생성되었고 c_0는 중도절단비율 (censoring proportion)이 15%가 되도록 조절되었다. 실험은 50회 반복되어 오차 측도로 평균제곱근오차 (root mean squared error; RMSE)와 평균절대값오차 (mean absolute error; MAE)를 사용하여 두 방법의 성능을 비교하였다.

$$RMSE_i = \sqrt{\frac{1}{20}\sum_{j=1}^{20}(\widehat{f}(x_{ij}) - f(x_{ij}))^2},\ MAE_i = \frac{1}{20}\sum_{j=1}^{20}\left|\widehat{f}(x_{ij}) - f(x_{ij})\right|,\ i = 1, \cdots, 6.$$

이 실험에서는 Gaussian커널, $K(x_1, x_2) = \exp\left(-(||x_1 - x_2||^2)/\sigma^2\right)$을 사용하였고, 각 반복에서의 벌칙상수와 커널모수의 최적값은 식 (3.13)의 일반화교차타당성함수를 이용하여 구하였다.

Table 4.1과 Table 4.2는 각각 두 가지 방법의 RMSE와 MAE의 평균을 나타낸다. 표에서 평균제곱근오차과 평균절대값오차의 감소율이 70% 이상 (대응되는 표본표준오차는 80% 이상) 임을 알 수 있고, 이로서 우리는 제안된 혼합효과 LS-SVM을 이용한 추정이 임의효과항을 고려하지 않는 일반적 LS-SVM을 이용한 추정보다 더 좋은 결과를 보여줌을 알 수 있다. Figure 4.1은 각 집단별로 추정된 회귀함수를 나타낸다. 여기서, 점은 중도절단되지 않은 관측시간, o는 중도절단된 관측시간을 나타내고, 실선 (solid line)은 실제 회귀함수, 파선 (dashed line)은 혼합효과 LS-SVM에 의해 추정된 회귀함수, 그리고 점선 (dotted line)은 일반적 LS-SVM에 의해 추정된 회귀함수를 나타낸다. 표와 그림에서 우리는 제안된 혼합효과 LS-SVM을 이용한 추정이 일반적 LS-SVM을 이용한 추정보다 일관되게 더 좋은 결과를 보여줌을 알 수 있다.

Table 4.1 Averages of RMSEs in each group (standard errors in parenthesis)

	group1	group2	group3	group4	group5	group6
LS-SVM	2.6732	2.5829	2.3916	2.9140	2.9338	2.9067
	(1.9733)	(1.8293)	(1.4954)	(2.2216)	(2.1961)	(1.9894)
mixed effects LS-SVM	0.6560	0.6547	0.6358	0.6278	0.6370	0.6317
	(0.3085)	(0.3231)	(0.2470)	(0.2696)	(0.2824)	(0.2658)
% of reduction	74.46	74.65	73.42	78.46	78.29	78.27
	(84.37)	(82.34)	(83.48)	(87.86)	(87.14)	(86.64)

Table 4.2 Averages of MAEs in each group (standard errors in parenthesis)

	group1	group2	group3	group4	group5	group6
LS-SVM	2.5757	2.4917	2.2992	2.8334	2.8463	2.8277
	(1.9859)	(1.8515)	(1.5164)	(2.2400)	(2.2204)	(2.0090)
mixed effects LS-SVM	0.5432	0.5385	0.5286	0.5130	0.5224	0.5276
	(0.2714)	(0.2764)	(0.2117)	(0.2317)	(0.2374)	(0.2305)
% of reduction	78.91	78.39	73.42	81.89	81.65	81.34
	(86.33)	(82.34)	(85.07)	(89.66)	(89.31)	(88.53)

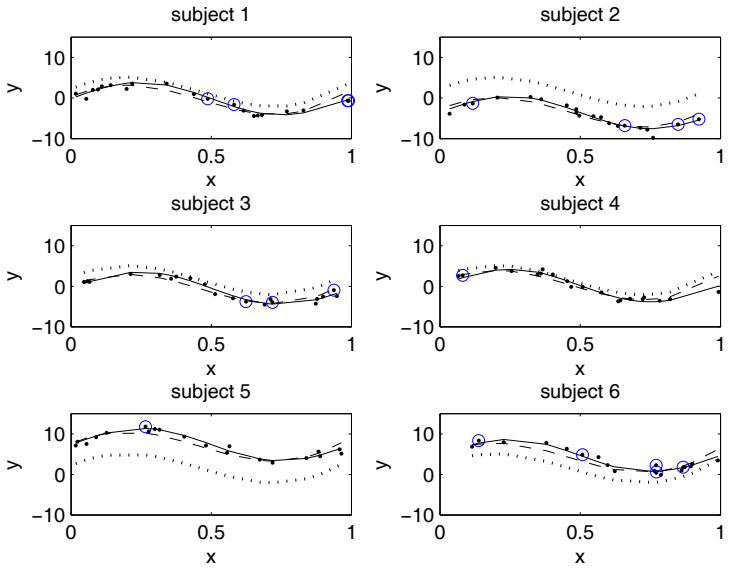

Figure 4.1 Estimates of regression functions in each group. Dashed line: Mixed LS-SVM estimate, Dotted line: LS-SVM estimate, Solid line: True regression function

5. 결론

본 논문에서는 각 집단별로 생존자료가 관측되는 경우 적용할 수 있는 혼합효과 LS-SVM이 제안되었다. 혼합효과 LS-SVM이 임의효과항을 고려하지 않는 일반적 LS-SVM보다 더 좋은 성능을 보임을 모의실험을 통하여 알 수 있었다. 더욱이 혼합효과 LS-SVM도 LS-SVM과 마찬가지로 벌칙상수와 커널모수의 최적의 값을 선택하는데 사용가능한 일반화교차타당성함수를 가지므로 집단별로 관측된 생존자료의 분석에 많이 활용될 것으로 예상된다.

참고문헌

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika*, **66**, 429-436.
Cox, D. R. (1972). Regression models과 life tables. *Journal of the Royal Statistical Society*, **34**, 187-202.
Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerishe Mathematik*, **31**, 377-390.

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press, Cambridge.

Davidian, M. and Giltinan, D. M. (1995). *Nonlinear models for repeated measurement data*, Chapman and Hall, London.

Green, P. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models*, Chapman and Hall, London.

Gunn, S. R. (1998). *Support vector machines for classification and regression*, Technical Report, Deptment of Electronics and Computer Science, Southampton University.

Harville, D. A. (1976). Extension of the Gauss-Markov theorem to include the estimation of random effects. *Annals of Statistics*, **4**, 384-395.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of American Statistical Association*, **72**, 320-340.

Hwang, C. and Shim, J. (2011). Cox proportional hazard model with L1 penalty. *Journal of the Korean Data & Information Society*, **22**, 613-618.

Jo, D. H., Shim, J. and Seok, K. H. (2010). Doubly penalized kernel method for heteroscedastic autoregressive data. *Journal of the Korean Data & Information Society*, **21**, 155-162.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The statistical analysis of failure time data*, John Wiley & Sons Inc., New York.

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

Kim, M., Park, H., Hwang, C and Shim, J. (2008). Claims reserving via kernel machine. *Journal of the Korean Data &Information Society*, **19**, 1419-1427.

Koul, H., Susarla, V. and Van Ryzin J. (1981). Regression analysis with randomly right censored data. *Annal of Statistics*, **9**, 1276-1288.

Laird, N. M. and Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics*, **56**, 89-97.

McCulloch, C, E. and Searle, S. R. (2000). *Generalized, linear, and mixed models*, John Wiley and Sons, New York.

Mercer, J. (1909). Functions of positive and negative type and their connection with theory of integral equations. *Philosophical Transactions of Royal Society A*, 415-446.

Miller, R. G. (1981). *Survival analysis*, Wiley, New York.

Moulton, L. H. and Dibley, M. J. (1997). Multivariate time-to-event models for studies of recurrent childhood diseases. *International Journal of Epidemiology*, **26**, 1334-1339.

Saunders, C., Gammerman, A. and Vovk, V. (1998). Ridge regression learning algorithm in dual variables. In *Proceedings of 15th International Conference on Machine Learning*, Madison, WI, 515-521.

Scholkopf, B. and Smola, A. (2002). *Learning with kernels-support vector machines, regularization, optimization and beyond*, Cambridge, MA, MIT Press.

Shim, J. and Lee, J. T. (2009). Kernel method for autoregressive data. *Journal of the Korean Data & Information Society*, **20**, 467-472.

Suykens, J. A. K. and Vandewalle, J. (1999). Least square support vector machine classifier. *Neural Processing Letters*, **9**, 293-300.

Vonesh, E. F. and Chinchilli, V. M. (1996). *Linear and nonlinear models for the analysis of repeated measurements*, Marcel Dekker, New York.

Vapnik, V. N. (1995). *The nature of statistical learning theory*, Springer, New York.

Vapnik, V. N. (1998). *Statistical learning theory*, John Wiley, New York.

Wahba, G. (1990). *Spline models for observational data*, BMS-NSF Regional Conference Series in Applied Mathematics, **59**, SIAM, Philadelphia.

Journal of the Korean Data & Information Science Society
2012, **23**(4), 739–748

http://dx.doi.org/10.7465/jkdi.2012.23.4.739
한국데이터정보과학회지

Mixed effects least squares support vector machine for survival data analysis[†]

Changha Hwang[1] · Jooyong Shim[2]

[1]Department of Statistics, Dankook University
[2]Department of Data Science, Inje University

Received 14 June 2012, revised 3 July 2012, accepted 9 July 2012

Abstract

In this paper we propose a mixed effects least squares support vector machine (LS-SVM) for the censored data which are observed from different groups. We use weights by which the randomly right censoring is taken into account in the nonlinear regression. The weights are formed with Kaplan-Meier estimates of censoring distribution. In the proposed model a random effects term representing inter-group variation is included. Furthermore generalized cross validation function is proposed for the selection of the optimal values of hyper-parameters. Experimental results are then presented which indicate the performance of the proposed LS-SVM by comparing with a standard LS-SVM for the censored data.

Keywords: Censoring distribution, generalized cross validation function, least squares support vector machine, mixed effects regression model, random effects, randomly right censored data.

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (2011-0027344).
[1] Professor, Department of Statistics, Dankook University, 126, Jukjeon-dong, Suji-gu, Yongin-si, Gyeonggi-do 448-701, Korea.
[2] Corresponding author: Adjunct professor, Department of Data Science, Institute of Statistical Information, Inje University, Obang-Dong, Kimhae 621-749, Korea. E-mail: ds1631@hanmail.net