단조 서포트벡터기계를 이용한 카플란-마이어 생존함수의 평활[†]

황창하¹ · 심주용²

¹단국대학교 정보통계학과 · ²인제대학교 데이터정보학과

접수 2012년 8월 30일, 수정 2012년 9월 17일, 게재확정 2012년 9월 30일

요약

서포트벡터기계는 분류 및 비선형 함수추정에서 유용하게 사용되고 있는 통계적 기법이다. 본 논문에서는 두 개의 입력변수와 회귀함수의 단조 관계를 이용하여 단조 서포트벡터기계를 제안하고, Kaplan-Meier의 방법에 의해서 생존함수의 추정값이 주어진 경우 제안된 방법을 이용하여 생존함수를 평활하는 방법 또한 제안한다. 모의실험에서는 실제 생존함수를 이용하여 Kaplan-Meier의 방법에 의한 생존함수의 추정값과의 성능을 비교함으로써 제안된 방법의 우수성을 보이기로 한다.

주요용어: 단조, 벌칙상수, 생존함수, 서포트벡터기계, 임의우측 중도절단자료, 커널기법, 커널모수, 평활.

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 발생하는 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.

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 SVM은

[†] 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

커널기법을 사용하여 선형 모델을 비선형 모델로 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 모델로서 기존의 선형 모델을 비선형 모델로 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 인공지능경망 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), Schölkopf와 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-SVM의 응용에 대한 자세한 내용은 참고문헌 Kim 등 (2008), Jo 등 (2010), Hwang과 Shim (2010, 2011)에 설명되어 있다. Kaplan-Meier 생존함수 추정값을 평활하는 방법에는 서포트벡터기계 및 평활 스플라인방법 (smoothing spline; Green 1994, Wahba, 1990)이 적용될 수 있으나 생존함수의 단조관계를 모형에 포함하지 않으므로 합리적인 평활을 할 수 없다.

본 논문은 다음과 같이 구성되어 있다. 2절에서는 회귀함수 추정을 위한 서포트벡터기계에 대해 간단히 기술하고 3절에서는 회귀함수가 단조관계를 가지고 종속변수가 두 개가 주어지는 경우 단조 서포트벡터기계를 제안한다. 4절에서는 단조 서포트벡터기계를 이용하여 Kaplan-Meier 생존함수 추정값을 평활하는 방법을 제안하고 5절에서는 모의실험에 대한 결과를 설명하고 6절에서는 결론을 기술한다.

2. 서포트벡터기계

입력벡터 $\mathbf{x}_i \in R^d$ 와 종속변수 $y_i \in R$ 로 이루어진 자료에서 우리는 다음과 같은 회귀모형을 가정한다.

$$m(\mathbf{x}_i) = \mathbf{w}'\phi(\mathbf{x}_i) + b, \quad i = 1, \dots, n, \quad (2.1)$$

여기서 b 는 편의항, $\phi(\mathbf{x}_i)$ 는 $d_f \times 1$ 비선형특징사상함수 (nonlinear feature mapping function), 그리고 $\mathbf{w}$ 는 $\phi(\mathbf{x})$ 에 대응되는 모수벡터이다. ϕ 에 대해 $\phi(\mathbf{x})'\phi(\mathbf{x}) = K(\mathbf{x}, \mathbf{x})$ 를 만족하는 커널함수가 존재한다. (Mercer, 1909). e-무감각 손실함수 (e-insensitive loss function)를 이용하여 Vapnik (1995, 1998)은 다음과 같은 목적함수의 최소화를 고려하는 서포트벡터기계를 제안하였다.

최소화 목적함수:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \quad (2.2)$$

제약조건:

$$\begin{aligned} y_i - (\mathbf{w}'\phi(\mathbf{x}_i) + b) &\leq e + \xi_i, \\ \mathbf{w}'\phi(\mathbf{x}_i) + b - y_i &\leq e + \xi_i^*, \xi_i, \xi_i^* \geq 0, \quad i = 1, \dots, n. \end{aligned} \quad (2.3)$$

식 (2.2)와 (2.3)을 이용하여 라그랑주 (Lagrange) 함수를 다음과 같이 만들 수 있다.

$$\begin{aligned} L = & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) - \sum_{i=1}^n \alpha_i (\xi_i - y_i^2 + (\mathbf{w}' \phi(\mathbf{x}_i) + b)) \\ & - \sum_{i=1}^n \eta_i \xi_i - \sum_{i=1}^n \alpha_i^* (\xi_i^* - (\mathbf{w}' \phi(\mathbf{x}_i) + b) + y_i^1) - \sum_i \eta_i^* \xi_i^* \\ & - \sum_{i=1}^{n-1} \beta_i (\mathbf{w}' (\phi(\mathbf{x}_i) - \phi(\mathbf{x}_{i+1}))). \end{aligned} \quad (2.4)$$

여기서 $\alpha_i^{(*)} \geq 0$, $\beta_i \geq 0$ 는 라그랑주배수이다. 식 (2.4)를 $(\mathbf{w}, b, \xi_i^{(*)})$ 에 관하여 편미분하면 다음과 같은 결과를 얻는다.

$$\mathbf{w} = \sum_{i=1}^n (\alpha_i - \alpha_i^*) \phi(\mathbf{x}_i), \sum_i (\alpha_i - \alpha_i^*) = 0, 0 \leq \alpha_i, \alpha_i^* \leq C. \quad (2.5)$$

식 (2.5)의 결과를 식(2.4)에 대입하고 커널함수를 이용하면 라그랑주배수 $(\alpha_i^{(*)}, \beta_i)$ 는 다음의 목적함수를 최소화함으로써 구해진다.

최소화 목적함수:

$$\frac{1}{2} \tilde{\boldsymbol{\alpha}}' H \tilde{\boldsymbol{\alpha}} - \sum_{i=1}^n y_i (\alpha_i - \alpha_i^*) + e \sum_{i=1}^n (\alpha_i + \alpha_i^*) \quad (2.6)$$

제약조건:

$$\sum_{i=1}^n (\alpha_i - \alpha_i^*) = 0, 0 \leq \alpha_i, \alpha_i^* \leq C, \beta_i \geq 0, i = 1, \dots, n.$$

여기서 $\tilde{\boldsymbol{\alpha}}' = (\boldsymbol{\alpha}', \boldsymbol{\alpha}^{*'})$ 이고 $H = \begin{pmatrix} K & -K \\ -K & K \end{pmatrix}$, $K = \{K(\mathbf{x}_i, \mathbf{x}_i)\}_{i=1}^n$ 는 $n \times n$ 행렬이다.

따라서 입력벡터 $\mathbf{x}$ 에 대응되는 회귀함수의 추정값은 다음과 같이 구해진다.

$$\hat{m}(\mathbf{x}) = K(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) + b, \quad (2.7)$$

그리고 b 는 KKT조건 (Karush-Kuhn-Tucker conditions; Kuhn과 Tucker, 1951)에 의해서 다음과 같이 구해진다.

$$b = \frac{1}{n_s} \sum_{i \in I_s} (y_i - K_i(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)), \quad (2.8)$$

여기서 n_s 는 $I_s = \{i = 1, \dots, n : 0 < \alpha_i < C, 0 < \alpha_i^* < C\}$ 의 크기이다.

3. 단조 서포트벡터기계

입력벡터 $\mathbf{x}_i \in R^d$ 와 종속변수 $(y_i^1, y_i^2) \in R \times R$ ($y_i^1 \leq y_i^2$)로 이루어진 자료에서 입력벡터 $\mathbf{x}_i$ 에 대해 회귀함수 $m(\mathbf{x}_i) = \mathbf{w}' \phi(\mathbf{x}_i) + b$ 가 다음과 같은 형태로 관련되어 있다고 가정한다.

$$m(\mathbf{x}_i) \geq m(\mathbf{x}_{i+1}), i = 1, \dots, n-1, \quad (3.1)$$

여기서 $\phi(\mathbf{x}_i)$ 는 $d_f \times 1$ 비선형특징사상함수, 그리고 $\mathbf{w}$ 는 $\phi(\mathbf{x}_i)$ 에 대응되는 모수벡터이다.

$m(\mathbf{x}_i) = \mathbf{w}'\phi(\mathbf{x}_i) + b$ 를 추정하기 위하여 우리는 다음과 같은 손실함수를 고려한다.

$$\sum_{i:m_i \leq y_i^1} (y_i^2 - m(\mathbf{x}_i)) + \sum_{i:m_i \geq y_i^2} (m(\mathbf{x}_i) - y_i^1). \quad (3.2)$$

우리는 식 (3.2)를 최소화하는 $(\mathbf{w}, b)$ 를 구하기 위하여 다음과 같은 0-무감각 서포트벡터기계 (0-insensitive support vector regression) 형태의 최적화 문제를 고려한다.

최소화 목적함수:

$$\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \quad (3.3)$$

제약조건:

$$\begin{aligned} y_i^2 - (\mathbf{w}'\phi(\mathbf{x}_i) + b) &\leq \xi_i, \\ \mathbf{w}'\phi(\mathbf{x}_i) + b - y_i^1 &\leq \xi_i^*, \xi_i, \xi_i^*, 0, i = 1, \dots, n, \\ \mathbf{w}'\phi(\mathbf{x}_i) &\geq \mathbf{w}'\phi(\mathbf{x}_{i+1}), i = 1, \dots, n-1. \end{aligned} \quad (3.4)$$

우리는 식 (3.3)과 (3.4)를 이용하여 라그랑주함수를 다음과 같이 만들 수 있다.

$$\begin{aligned} L = & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) - \sum_{i=1}^n \alpha_i (\xi_i - y_i^2 + (\mathbf{w}'\phi(\mathbf{x}_i) + b)) \\ & - \sum_{i=1}^n \eta_i \xi_i - \sum_{i=1}^n \alpha_i^* (\xi_i^* - (\mathbf{w}'\phi(\mathbf{x}_i) + b) + y_i^1) \\ & - \sum_i \eta_i^* \xi_i^* - \sum_{i=1}^{n-1} \beta_i (\mathbf{w}'(\phi(\mathbf{x}_i) - \phi(\mathbf{x}_{i+1}))). \end{aligned} \quad (3.5)$$

여기서 $\alpha_i^{(*)} \geq 0$, $\beta_i \geq 0$ 는 라그랑주배수이다. 식 (3.5)를 $(\mathbf{w}, b, \xi_i^{(*)})$ 에 관하여 편미분하면 다음과 같은 결과를 얻는다.

$$\mathbf{w} = \sum_{i=1}^n (\alpha_i - \alpha_i^*) \phi(\mathbf{x}_i) + \sum_{i=1}^{n-1} \beta_i (\phi(\mathbf{x}_i) - \phi(\mathbf{x}_{i+1})), \sum_i (\alpha_i - \alpha_i^*) = 0, 0 \leq \alpha_i, \alpha_i^* \leq C. \quad (3.6)$$

식 (3.6)의 결과를 식(3.5)에 대입하고 커널함수를 이용하면 라그랑주배수 $(\alpha_i^{(*)}, \beta_i)$ 는 다음의 목적함수를 최소화함으로써 구해진다.

최소화 목적함수:

$$\frac{1}{2} \tilde{\alpha}' H \tilde{\alpha} - \sum_{i=1}^n (y_i^2 \alpha_i - y_i^1 \alpha_i^*) \quad (3.7)$$

제약조건:

$$\sum_{i=1}^n (\alpha_i - \alpha_i^*) = 0, 0 \leq \alpha_i, \alpha_i^* \leq C, \beta_i \geq 0.$$

여기서 $\tilde{\alpha}' = (\alpha', \alpha^{*'}, \beta')$ 이고 $H = \begin{pmatrix} K & -K & K_{.1} - K_{.2} \\ -K & K & K_{.2} - K_{.1} \\ K'_{.1} - K'_{.2} & K'_{.2} - K'_{.1} & K_{11} - K_{12} - K_{21} + K_{22} \end{pmatrix}$, $K = \{K(\mathbf{x}_i, \mathbf{x}_i)\}_{i=1}^n$ 는 $n \times n$ 행렬, $K_{.1} = \{K(\mathbf{x}_i, \mathbf{x}_j)\}_{i=1, j=1}^{n, n-1}$ 는 $n \times (n-1)$ 행렬, $K_{.2} = \{K(\mathbf{x}_i, \mathbf{x}_j)\}_{i=1, j=2}^{n, n}$ 는 $n \times (n-1)$ 행렬, $K_{11} = \{K(\mathbf{x}_i, \mathbf{x}_i)\}_{i=1}^{n-1}$ 는 $(n-1) \times (n-1)$ 행렬, $K_{22} = \{K(\mathbf{x}_i, \mathbf{x}_i)\}_{i=2}^n$ 는 $(n-1) \times (n-1)$ 행렬, $K_{12} = \{K(\mathbf{x}_i, \mathbf{x}_j)\}_{i=1, j=2}^{n-1, n}$ 는 $(n-1) \times (n-1)$ 행렬, $K_{21} = \{K(\mathbf{x}_i, \mathbf{x}_j)\}_{i=2, j=1}^{n, n-1}$ 는 $(n-1) \times (n-1)$ 행렬이다.

따라서 입력벡터 $\mathbf{x}$ 에 대응되는 회귀함수의 추정값은 다음과 같이 구해진다.

$$\hat{m}(\mathbf{x}) = K(\alpha - \alpha^*) + (K_{.1} - K_{.2})\beta + b, \quad (3.8)$$

그리고 b 는 KKT조건에 의해서 다음과 같이 구해진다.

$$b = \frac{1}{n_s} \sum_{i \in I_s} (y_i - K_i(\alpha - \alpha^*) - (K_{1,i} - K_{2,i})\beta),$$

여기서 n_s 는 $I_s = \{i = 1, \dots, n : 0 < \alpha_i < C, 0 < \alpha_i^* < C\}$ 의 크기이다.

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

$$CV(\theta) = \frac{1}{n} \left(\sum_{i: m_i \leq y_i^1} (y_i^2 - \hat{m}_\theta^{(-i)}(\mathbf{x}_i)) + \sum_{i: m_i \geq y_i^2} (\hat{m}_\theta^{(-i)}(\mathbf{x}_i) - y_i^1) \right). \quad (3.9)$$

여기서 θ 는 벌칙상수와 커널모수로 이루어진 벡터이고, $\hat{m}_\theta^{(-i)}(\mathbf{x}_i)$ 는 전체자료 중 i 번째 자료를 제외한 나머지 자료들을 이용하여 구한 $m_\theta(\mathbf{x}_i)$ 의 추정값이다.

4. Kaplan-Meier 생존함수의 평활

주어진 생존자료에서 실제로, 우리들은 종속변수인 생존시간 t 를 관측할 수는 없고 관측변수 $\min(t, c)$ 와 $I(t \leq c)$ 를 관측할 수 있다. 여기서, c 는 t 에 대응하는 중도절단변수 (censoring variable)이다. 중도절단변수들은 임의우측 중도절단 (randomly right censored) 되었고 서로 독립으로 가정된다. 대부분의 경우에 생존함수 $S(t)$ 는 알려지지 않으므로 주로 Kaplan-Meier (1958)의 방법으로 다음과 같이 추정된다.

$$p_i = 1 - \frac{d_i}{n_i}, \quad S(t_i) = S(t_{i-1}) \times p_i = \prod_{j=1}^i \left(1 - \frac{d_j}{n_j}\right), \quad i = 1, \dots, m,$$

여기서 n_i 는 i 번째 구간의 시작점 직전의 생존자 수 (number of individuals at risk)이고 d_i 는 구간에서의 사망자수, i 번째 구간이란 i 번째 사망시점에서 다음 사망시점 직전까지의 구간이다. 그리고 $S(t_i)$ 는 i 번째 구간에서 생존할 확률을 의미한다. 편의상 $S(t_i)$ 를 S_i 로 표기하면 $\mathbf{s}^1 = (S_2, \dots, S_m, 0)'$ 와 $\mathbf{s}^2 = (1, S_1, \dots, S_{m-1})'$ 는 시점 $(z_1, z_2, \dots, z_m)'$ 에서 구해진 Kaplan-Meier 생존함수의 하위 및 상위 경계가 된다. 따라서 우리는 $\mathbf{z} = (z_1, \dots, z_m)'$ 를 입력벡터로, $(\mathbf{s}^1, \mathbf{s}^2)$ 를 종속변수벡터들로 사용하여 생존함수의 Kaplan-Meier 추정값의 평활값을 단조 서포트벡터기계를 이용하여 구할 수 있다.

5. 실험 및 결과

이절에서는 모의자료와 실제자료를 이용하여, 제안된 단조 서포트벡터기계를 이용한 생존함수의 Kaplan-Meier 추정값 평활의 유한 표본에 대한 성능을 설명한다. 그리고 제안된 방법은 MatLab을 이용하여 구현하였다우리는 생존시간, 중도절단시간 및 관측시간을 다음과 같이 모의자료를 생성하였다.

$$t_i \sim \exp(1), c_i \sim \exp(1) + c_0, z_i = \min(t_i, c_i), \delta_i = I(t_i \leq c_i), i = 1, \dots, 50,$$

여기서 c_0 는 중도절단 비율 (censoring proportion)이 10%와 30%가 되도록 조절되었다. 오차 측도로서 평균제곱근오차 (root mean squared error; RMSE)와 Kolmogorov-Smirnov 통계량 (Chakravarti 등, 1967)을 사용하여 두 방법의 성능을 비교하였다.

$$RMSE_k = \sqrt{\frac{1}{50} \sum_{i=1}^{50} (\hat{S}_k(z_i) - S(z_i))^2}, D_k = \sup_{z_i} |\hat{S}_k(z_i) - S(z_i)|, k = 1, 2,$$

여기서 $k = 1$ 은 Kaplan-Meier 생존함수 추정, $k = 2$ 는 단조 서포트벡터기계를 이용한 평활값이고, $S(z)$ 는 지수분포(1)을 이용한 생존함수, $\hat{S}_1(z)$ 은 생존함수의 Kaplan-Meier 추정값이고 $\hat{S}_2(z)$ 은 Kaplan-Meier 추정값의 단조 서포트벡터기계를 이용한 평활값을 나타낸다. 실험은 200회 반복되었고, 이 실험에서는 Gaussian 커널, $K(z_1, z_2) = \exp(-(|z_1 - z_2|)^2 / \sigma^2)$ 을 사용하였고, 각 반복에서의 별척상수와 커널모수의 최적값은 식 (3.9)의 LOO 교차타당성함수를 이용하여 구하였다.

Table 5.1은 두 가지 방법의 평균제곱근오차와 KS-통계량의 평균 및 표본표준오차를 나타낸다. 표에서 우리는 제안된 단조 서포트벡터기계를 이용한 평활값이 Kaplan-Meier 생존함수 추정보다 더 작은 오차를 보여줌을 알 수 있다. Figure 5.1은 두 가지 방법의 평균제곱근오차와 KS-통계량의 상자그림을 각각 나타낸다. Figure 5.2는 200개 표본 중 하나의 표본에서 생존함수를 나타낸다. 여기서, 실선은 지수분포(1)을 이용하여 구한 실제 생존함수, 계단-실선은 Kaplan-Meier 생존함수 추정값 그리고 점선은 Kaplan-Meier 생존함수 추정값의 단조 서포트벡터기계를 이용한 평활값을 나타낸다. 표와 그림에서 우리는 제안된 단조 서포트벡터기계를 이용한 평활값이 Kaplan-Meier 생존함수 추정보다 주어진 표본에서 일관되게 더 좋은 결과를 보여줌을 알 수 있다.

Stablein 등 (1981)이 사용한 자료는 화학요법으로 치료받은 45명의 위암환자와 화학-방사능요법으로 치료받은 45명의 위암환자의 관측시간으로 구성되어 있다. Figure 5.3은 주어진 자료의 Kaplan-Meier 추정값과 Kaplan-Meier 생존함수 추정값의 단조 서포트벡터기계를 이용한 평활값을 나타낸다. 우리는 그림에서 200일에서 300일 사이의 관측시간에서 화학요법 집단의 생존함수가 화학-방사능요법의 생존함수 더 크다는 것을 쉽게 알 수 있다. 그리고 약 990일에 두 집단의 생존함수가 교차함을 알 수 있다. 제안된 방법에 의한 250일의 화학요법 집단과 화학-방사능요법의 생존함수 추정값은 각각 0.8444와 0.5142이고, 1000일의 화학요법 집단과 화학-방사능요법의 생존함수 추정값은 각각 0.2134와 0.2169이다.

Table 5.1 Averages of RMSEs and KS-statistics (standard errors in parenthesis)

	censoring proportion	RMSE	KS-statistic
Kaplan-Meier	10%	0.0568 (0.0016)	0.1154 (0.0025)
	30%	0.0803 (0.0023)	0.1758 (0.0044)
proposed	10%	0.0535 (0.0016)	0.0956 (0.0024)
	30%	0.0786 (0.0022)	0.1626 (0.0045)

Figure 5.1 Boxplots of RMSEs (left) and KS-statistics (right) for 10% censoring proportion (upper) and 30% censoring proportion (lower). 1: Kaplan-Meier, 2: proposed

Figure 5.2 Estimates of survival functions for 10% censoring proportion (left) and 30% censoring proportion (right). solid line: true, step-solid line: Kaplan-Meier, dotted line: proposed

Figure 5.3 Estimates of survival functions for chemotherapy treatment group (left) and 30% censoring proportion (right). solid line: true, step-solid line: Kaplan-Meier, dotted line: proposed

6. 결론

본 논문에서는 단조 서포트벡터기계를 제안하였고, 생존자료 분석에서 생존함수의 Kaplan-Meier 추정값이 구해진 경우 이것을 단조 서포트벡터기계를 이용하여 평활하는 법도 제안하였다. 실험을 통하여 제안된 방법이 만족할만한 결과를 산출함을 보였다. 그리고 제약조건을 조정하면 회귀함수가 단조증가인 경우에도 적용이 가능한 단조 서포트벡터기계도 구해질 수 있다. 향후 단조 서포트벡터기계 모형선택 방법으로 LOO 교차타당성함수보다 더 효율적인 방법의 개발이 필요하다.

참고문헌

- Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika*, **66**, 429-436.
- Chakravarti, I. M., Laha, R. G. and Roy, J. (1967). *Handbook of methods of applied statistics*, John Wiley and Sons, New York.
- Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society*, **34**, 187-202.
- Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press, Cambridge.
- Green, P. J. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models*, Chapman & Hall, London.
- Gunn, S. R. (1998). *Support vector machines for classification and regression*, Technical Report, Department of Electronics and Computer Science, Southampton University.
- Hwang, C. and Shim, J. (2010). Semiparametric support vector machine for accelerated failure time model. *Journal of the Korean Data & Information Science Society*, **21**, 467-477.
- 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 Science 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.
- Kuhn, H. W. and Tucker, A. W. (1951). Nonlinear programming. In *Proceedings of 2nd Berkeley Symposium*, University of California Press, Berkeley, 481-491.

- Kim, M., Park, H., Hwang, C. and Shim, J. (2008). Claims reserving via kernel machine. *Journal of the Korean Data & Information Science Society*, **19**, 1419-1427.
- 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.
- Schölkopf, B. and Smola, A. (2002). *Learning with kernels-support vector machines, regularization, optimization and beyond*, MIT Press, Cambridge, MA.
- Stablein, D. M., Carter, W. H. and Novak, J. W. (1981). Analysis of survival data with nonproportional hazard functions. *Controlled Clinical Trials*, **2**, 149-159.
- Suykens, J. A. K. and Vandewalle, J. (1999). Least square support vector machine classifier. *Neural Processing Letters*, **9**, 293-300.
- 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*, SIAM, Philadelphia.

Smoothing Kaplan-Meier estimate using monotone support vector regression[†]

Changha Hwang¹ · Jooyong Shim²

¹Department of Statistics, Dankook University

²Department of Data Science, Inje University

Received 30 August 2012, revised 17 September 2012, accepted 30 September 2012

Abstract

Support vector machine is known to be the very useful statistical method in classification and nonlinear function estimation. In this paper we propose a monotone support vector regression (SVR) for the estimation of monotonically decreasing function. The proposed monotone SVR is applied to smooth the Kaplan-Meier estimate of survival function. Experimental results are then presented which indicate the performance of the proposed monotone SVR using survival functions obtained by exponential distribution.

Keywords: Kernel parameter, kernel technique, monotonicity, randomly right censored data, smoothing, support vector machine, survival function.

[†] 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. Gyeonggi-do 448-701, Korea

² Corresponding author: Adjunct professor, Department of Data Science, Institute of Statistical Information, Inje University, Kimhae 621-749, Korea. E-mail: ds1631@hanmail.net