Journal of the Korean Data & Information Science Society
2016, **27**(4), 993–1000

http://dx.doi.org/10.7465/jkdi.2016.27.4.993
한국데이터정보과학회지

준지도 학습의 모수 선택에 관한 연구[†]

석경하[1]

[1]인제대학교 통계학과

접수 2016년 5월 16일, 수정 2016년 6월 16일, 게재확정 2016년 7월 22일

요 약

반응 값이 없는 자료를 지도학습 (supervised learning)에 사용하는 준지도 학습 (semi-supervised learning)은 분류에 더 많은 관심을 갖는다. 본 연구는 준지도학습을 회귀분석에 적용하는 준지도 회귀함수 추정법을 제안한다. 제안된 방법은 기존의 방법과 형태는 같지만 반응 값이 있는 자료와 없는 자료의 주변분포를 다르게 가정하고, 서로 다른 평활계수를 사용하는 등 좀 더 일반화된 형태를 가진다. 제안된 추정법의 점근분포를 계산하고 점근평균제곱오차를 최소화하는 최적의 평활계수가 가지는 조건을 찾는다. 설명변수의 주변분포에 대한 추정이 잘 이루어지고, 반응 값이 있는 자료와 없는 자료의 크기에 대한 조건을 적절하게 통제할 수 있고, 그리고 평활계수가 적절하게 선택될 수 있다면 라벨없는 자료가 회귀분석에서도 도움을 줄 수 있음을 보인다. 그리고 준지도 분류에서 사용하는 것처럼 반응 값이 없는 자료의 초기추정은 작은 값을 가지는 평활계수를 사용하여 과적합 (overfitting)되도록 하는 것이 좋음을 증명한다.

주요용어: 수렴율, 점근평균오차, 준지도 회귀분석, 커널회귀분석, 평활계수.

1. 서론

기계학습 (machine learning)과 통계적 학습 (statistical learning)의 목적은 예측 (prediction)과 군집분석 (clustering)이다. 반응변수 (response variable, label)를 예측하는 회귀분석과 분류 (classification)는 반응변수를 추구한다는 관점에서 지도학습 (supervised learning)으로, 설명변수만 사용하는 군집분석은 자율학습 (unsupervised learning)으로 구분된다.

영상인식 (image classification), 생명정보 (bioinformatics), 음성인식 (speech recognition), 문자분류 (text categorization) 그리고 웹분류 (web categorization) 분야에서 발생하는 자료 중에는 반응 값을 만드는 것에 시간이나 경비를 많이 지불해야 하는 경우가 있다. 그러한 이유로 반응 값이 있는 자료 (라벨자료, labeled data)보다 반응값이 없는 자료 (언라벨자료, unlabeled data)가 훨씬 더 풍부할 수 있다.

언라벨자료를 예측에 사용하는 준지도학습 (semi-supervised learning)이 개발되면서 언라벨자료의 가치가 다시 평가되었다 (Zhu, 2005; Chapelle 등, 2006; Xu 등, 2010). 준지도학습은 언라벨자료도 지도학습에 사용할 수 있는 방법에 관한 것인데 개발된 방법은 자가훈련 (self training)과 상호훈련 (co-training)으로 나누어 진다. 대표적인 연구로는, 준지도학습에 관한 이론과 기초를 제공한 Niyogi (2008)과 준지도학습이 지도학습보다 더 좋은 수행력을 가질 수 있음을 보인 Zhu 와 Goldberg (2009)

[†] 이 논문은 2011년도 정부 (교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임 (2011-0009705).

[1] (50834) 경남 김해시 인제로 197, 인제대학교 통계학과, 통계정보연구소, 교수. E-mail: statskh@inje.ac.kr

등이 있다. 최근에는 Liu 등 (2014)과 Wei-Yu 등 (2015)이 각각 최소제곱 서포터벡터와 최소제곱법을 이용하여 영상인식에 적용하는 준지도 학습을 제안하였다. 이와 같이 준지도학습은 분류에 많이 사용되면서 최근에 많은 관심을 받고 있는 딥 러닝 (deep learning)과 빅데이터에서도 사용되고 있다.

언라벨자료가 회귀분석에서도 유용하게 사용될 수 있음은 Belkin 등 (2006)이 보였다. 그들은 분류에서 사용되는 그래프기반 (graph based) 준지도학습이 회귀분석에서도 사용될 수 있음을 보여주었다. Wang 등 (2006) 과 Seok (2013)은 각각 커널회귀모형 (kernel regression model)과 커널능형회귀모형 (kernel ridge regression model)을 기반으로 준지도회귀모형을 개발하였다. 그리고 Cortes와 Mohri (2007)도 커널회귀모형을 기반으로 준지도회귀모형을 개발하였는데 제안된 방법은 오직 라벨자료의 라벨 추정에만 관심 있는 전환적 (transductive) 방법인데 실험을 통해 큰 자료에서도 수행결과가 좋다는 것을 보여주었다. Lafferty와 Wasseman (2008)은 그래프 라플라시안 (Laplacian)을 사용하는 정칙성 (regularization) 기반 준지도회귀모형이 커널회귀모형보다 더 빠른 최소최대 수렴속도 (minimax convergence rate)를 가질 수는 없지만 평활성 (smoothness)에 대한 가정을 완화한다면 더 나은 준지도모형을 만들 수 있음을 언급하였다. Xu 등 (2011)은 최소제곱 서포트벡터 회귀모형 (least square support vector regression; Suykens 등, 2002)을 이용한 준지도 최소제곱 서포트벡터 회귀모형을 개발하였고 제안된 방법의 효율성을 입증하였다. 최근에 Seok (2015)은 준지도학습을 분위수 회귀모형에 응용한 준지도 서포트벡터 분위수 회귀모형을 개발하였다. 개발된 준지도 회귀모형의 점근 분포를 밝혀 커널회귀모형보다 더 빠른 수렴속도를 가질 수 있는 조건을 찾았다. 그리고 모의실험을 통해 타당성을 보였다.

분류를 위한 준지도 방법과는 달리 회귀분석을 위한 연구는 상당히 제한적이다. 또한 회귀분석을 위해 이루어진 연구도 제안된 준지도 방법의 우수성과 효율성을 입증하고 있지만 대개의 연구에서 교차타당성 (cross validation)과 같은 경험적인 방법에 의존하여 모수를 선택한다.

NWR은 에 의해 제안된 것으로 비모수 회귀모형의 기본라고 할 수 있다. 본 연구에서는 NWR에 기반한 준지도회귀모형 (semi-supervised NWR; SSNWR)을 제안한다. 그리고 제안된 모형의 점근분포를 계산하여, 이를 바탕으로 점근평균제곱오차 (asymptotic mean integrated squared error; AMISE)을 최소화 할 수 있는 평활계수를 알아본다. 그리고 AMISE가 최소가 되기 위해서 언라벨자료의 크기와 라벨자료의 크기 사이에 존재하는 관계를 알아본다.

준지도 회귀분석은 언라벨자료의 반응값을 추정하는 초기추정 (pilot estimation)과 언라벨자료와 라벨자료를 결합하여 최종 예측하는 두 단계로 구성된다. 일반적으로 언라벨자료의 크기가 더 큰 환경에서는 초기추정의 결과가 최종 결과에 많은 영향을 끼치므로 좋은 초기추정이 요구된다. 본 연구에서는 언라벨자료의 크기에 따른 평활계수선택에 대한 기준을 마련하였다. 커널추정법에서 사용되는 평활계수는 일반적으로 $O(n^{-1/5})$의 값을 가지고 이 때의 AMISE 는 $O(n^{-4/5})$가 된다. 그렇지만 본 연구에서는 AMISE가 최적수렴율을 가지기 위해서 초기추정에 사용되는 평활계수의 크기는 $o(n^{-1/5})$이 됨을 밝혔다. 즉 초기추정은 과적합되는 것이 더 좋다는 것을 증명하였다.

본 논문의 구성은 다음과 같다. 2절에서는 NWR을 소개하고 3절에서는 SSNWR을 제안한다. 그리고 4절에서는 제안된 추정량의 점근분포 및 평활계수선택에 대해 알아본다. 그리고 마지막으로 5절에서는 결론을 제시한다.

2. Nadaraya-Watson 회귀모형

NWR은 커널회귀모형과 평활법 (smoothing)의 기초가 되는 비모수적 회귀함수추정법이다. (X_1, Y_1),

$\cdots, (X_n, Y_n)$이 다음과 같은 회귀함수의 관계를 가지는 이변량함수라고 하자.

$$Y_i = m(X_i) + \epsilon_i,\ E(\epsilon_i) = 0,\ Var(\epsilon_i) = \sigma^2(x_i),\ i = 1, \cdots, n.$$

주어진 자료 $(X_1, Y_1), \cdots, (X_n, Y_n)$를 활용하여 회귀함수 $m(x) = E(Y|X = x)$를 추정하는 것이 목표다. X의 주변확률밀도함수와 $Y|X$의 조건부 확률밀도함수 그리고, X와 Y의 결합확률밀도함수를 단순히 $f(x), f(y|x)$그리고 $f(x, y)$로 각각 표기한다. 그러면 $m(x)$를 다음과 같이 $f(x, y)$를 이용해서 표현할 수 있다.

$$m(x) = E[Y|X = x] = \int yf(y|x)dy = \frac{\int yf(x, y)dy}{\int f(x, y)dy} \tag{2.1}$$

(2.1)의 분자와 분모를 따로 분리해서 추정하는 방법을 강구한다. 먼저 커널 밀도함수추정 (kernel density estimation)을 이용하여 $f(x, y)$를 추정한다.

$$\begin{aligned} \widehat{f}(x, y) &= \frac{1}{nh_xh_y} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_x}\right) K\left(\frac{y - y_i}{h_y}\right) \\ &= \frac{1}{n} \sum_{i=1}^{n} K_{h_x}(x - x_i)K_{h_y}(y - y_i). \end{aligned} \tag{2.2}$$

여기에서 $K_h(x) = \frac{1}{h}K(\frac{x}{h})$는 커널함수, h_x와 h_y는 평활계수다. 많이 사용되는 대표적인 커널함수는 다음과 같다.

$$\begin{aligned} &\text{가우시안 커널} : K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right),\ -\infty < x < \infty \\ &\text{Epanechnikov 커널} : K(x) = \frac{3}{4}(1 - u^2),\ |x| \le 1 \\ &\text{Triweight 커널} : K(x) = \frac{35}{32}(1 - u^2)^3,\ |x| \le 1. \end{aligned}$$

식 (2.1) 의 분자에 (2.2)를 대입하면 다음과 같다.

$$\int y\widehat{f}(x, y)dy = \frac{1}{n} \int y \sum_{i=1}^{n} K_{h_x}(x - x_i)K_{h_y}(y - y_i)dy$$

여기에서 $\int yK_{h_y}(y - y_i)dy = y_i$임을 이용하여 분자의 추정식을 다음과 같이 표현할 수 있다.

$$\int y\widehat{f}(x, y)dy = \frac{1}{n} \sum_{i=1}^{n} K_{h_x}(x - x_i)y_i$$

식 (2.1)의 분모는 $\int \widehat{f}(x, y)dy = \widehat{f}(x)$가 됨을 이용하면 NWR은 다음과 같이 주어진다.

$$\widehat{m}_{NW}(x) = \frac{\sum_{i=1}^{n} Y_iK_h(X_i - x)}{\sum_{i=1}^{n} K_h(X_i - x)}.$$

이 추정량은 Y_i에 대해 선형이기 때문에 선형 평활법 (linear smoother)으로 불리기도 한다.

3. 준지도 NWR

이 절에서는 준지도 NWR (semi-supervised Nadaraya-Watson regression estimator; SSNWR)을 소개하고자 한다. 먼저 반응변수 Y가 있는 라벨자료와 없는 언라벨자료를 각각 $D_l = \{(X_i, Y_i)\}_{i=1}^{l}$, $D_u = \{X_j\}_{j=l+1}^{n}$로 표기한다. 여기에서 l과 u는 라벨자료와 언라벨자료의 크기를 나타내고 $n = l + u$는 전체자료의 크기를 나타낸다. Wang 등 (2006)과 Seok (2012)은 다음과 같이 NWR를 이용한 준지도 추정량을 제안하였다.

$$m_{ss}(x) = \frac{S_x(D_l, YK_h) + \lambda S_x(D_u, YK_h)}{S_x(D_l, K_h) + \lambda S_x(D_u, K_h)}.$$

여기에서 $S_x(D_{b,}YK_h) = \frac{1}{b}\sum_{D_b} Y_i K_h(X_i - x)$, $S_x(D_{b,}K_h) = \frac{1}{b}\sum_{D_b} K_h(X_i - x)$, $D_b = D_l$혹은$D_b = D_u$, $b = |D_b|$그리고 λ는 라벨자료와 언라벨자료의 비중을 조절하는 가중인자다. 여기에서 사용된 평활계수h는D_l과D_u에서 같은 값을 사용했다. 언라벨자료의 반응변수 $Y_{j,}j = l+1, \cdots, n$에 대한 초기추정치는 다음과 같이 NWR을 사용하였다.

$$Y_j = \frac{S_{X_j}(D_l, YK_{h_0})}{S_{X_j}(D_l, K_{h_0})},\ j = l+1, \cdots, n. \tag{3.1}$$

여기에서 사용한 평활계수는 $h_0 = O(l^{-1/5})$다. 초기추정치를 이용하여 언라벨자료를 $D_u = \{(X_j, Y_j)\}_{j=l+1}^{n}$로 나타낸다. 간편하게 표기하기 위해 $\widehat{Y}_j$를 Y_j로 표현하였다. Wang 등 (2006)은 제안된 방법의 우수성을 경험적으로 보였고, Seok (2012)은 가중인자 λ를 적절히 조정하면 커널회귀모형보다 더 빠른 수렴속도를 가질 수 있음을 보였고 경험적인 방법을 이용한 실험을 통해 입증하였다.

본 연구에서는 좀 더 일반화된 SSNWR을 소개하고, 추정량의 점근분포를 알아본다. 그리고 AMISE를 최소화할 수 있는 조건을 찾는다. 먼저 D_l에서의X의 주변분포 f_l과D_u에서의 주변분포 f_u가 f_t로 같다는 가정에서 고려된 SSNWR1을 소개한다.

$$\widehat{m}_{ss1}(x) = \frac{S_x(D_t, YK_{h_t})}{S_x(D_t, K_{h_t})}. \tag{3.2}$$

여기에서 $D_t = D_l \cup D_u$이고 $Y_j, j = l+1, \cdots, n$는 초기추정량을 (3.2) 과 같은 것을 사용하였다. 그렇지만 여기에서 사용된 평활계수 h_t은 초기추정량에 사용된 h_0와 다르다. 두 번째로 소개되는 추정량 SSNWR2는 D_l에서의 X의 주변분포 f_l과 D_u에서의 주변분포 f_u가 다르다는 가정에서 고려되었다.

$$\widehat{m}_{ss2}(x) = \frac{S_x(D_l, YK_{h_l}) + S_x(D_u, YK_{h_u})}{S_x(D_l, K_{h_l}) + S_x(D_u, K_{h_u})}.$$

라벨자료에서 사용되는 평활계수 h_l, 언라벨자료에서 사용되는 평활계수 h_u그리고 초기 추정량에 사용되는 평활계수 h_0는 각각 다른 값을 사용한다. 다음 절에서는 제안된 추정량들의 점근분포를 알아본다.

4. SSNWR의 점근분포

이 절에서는 제안된 추정량들의 점근분포를 알아보고 AMISE의 수렴율을 최대로 하는 평활계수와 표본크기 l과 u의 관계를 알아본다.

4.1. 추정량의 점근분포

SSNWR1 과 SSNWR2의 점근분포를 알아보기 위해 먼저 회귀함수 m, 독립변수의 주변확률 밀도함수 그리고 커널함수에 대해 다음과 같은 전형적인 조건을 가정한다.

1. 회귀함수 $m(x)$는 유한이고 연속인 2차 미분을 가진다.
2. $var(Y|x) = \sigma^2(x)$는 유한이고 연속이다.
3. 라벨자료, 언라벨자료 그리고 전체자료의 독립변수에 대한 주변 확률밀도함수 f_l, f_u 그리고 f는 연속, 유한이고 연속인 2차 미분을 가진다.
4. 커널함수 K는 유한이고 연속인 확률밀도함수이고 $\int_{-\infty}^{\infty} xK(x)dx=0$, $\int_{-\infty}^{\infty} x^4K(x)dx<\infty$이다.

본 논문에서는 다음과 같은 표현을 사용한다.

$$\mu_i = \int_{-\infty}^{\infty} u^i K(u)du,\ \nu_i = \int_{-\infty}^{\infty} u^i K^2(u)du,\ B_v(x) = \frac{1}{2}m"(x) + \frac{m'(x)f_v'(x)}{f_v(x)},\ v = l, u, t.$$

정리 4.1 (Wasserman, 2006) 조건 1-4를 만족하고, $h_l \to 0$이고 $lh_l \to \infty$이라면 $\widehat{m}_{NW}(x)$는 다음과 같은 점근분포를 가진다.

$$\sqrt{lh_l}(\widehat{m}_{NW}(x) - m(x) - h_l^2\mu_2 B_l(x)) \to N\left(0, \frac{\sigma^2(x)\nu_0}{f_l(x)}\right)$$

정리 4.2 조건 1-4를 만족하고, h_0, $h_t \to 0$, $lh_0 \to \infty$ 그리고 $nh_t \to \infty$이라면 $\widehat{m}_{ss1}(x)$는 다음과 같은 점근분포를 가진다.

$$\sqrt{n_t h_t}(\widehat{m}_{ss1}(x) - m(x) - h_t^2\mu_2 B_t(x) + \frac{n_u}{n_t}h_0^2\mu_2 B_l(x)) \to N\left(0, \frac{\sigma^2(x)\nu_0}{f(x)}\right).$$

정리 4.3 조건 1-4를 만족하고, h_0, h_l, $h_u \to 0$이고 lh_0, lh_l, $uh_u \to \infty$이라면 $\widehat{m}_{ss2}(x)$는 다음과 같은 점근분포를 가진다.

$$\sqrt{lh_l + uh_u}(\widehat{m}_{ss2}(x) - m(x) - c) \to N(0, \sigma^2(x)\nu_0),$$
$$c = \frac{lh_l^3\mu_2 f_l(x)B_l(x) + uh_u^3\mu_2 f_u(x)B_u(x) + uh_u h_0^2\mu_2 f_u(x)B_l(x)}{lh_l f_l(x) + uh_u f_u(x)}.$$

증명: 위 정리는 다음의 사실들을 이용하여 쉽게 증명할 수 있다.

$$\begin{aligned}
E\{S_x(D_l, K_h)\} &= \frac{1}{l}E\left\{(\sum K_h(X_i - x)\right\} \\
&= f(x) + \frac{1}{2}h^2 f"(x)\mu_2 + o(h^2), \\
E\{S_x(D_l, YK_h)\} &= \frac{1}{l}E\left\{(\sum Y_i K_h(X_i - x)\right\} \\
&= m(x)f(x) + h^2\mu_2\left\{m'(x)f'(x) + (m"(x)f(x) + m(x)f"(x))/2\right\} + o(h^2), \\
Var(S_x(D_l, K_h)) &= \frac{f(x)\nu_0}{nl} + o\left(\frac{1}{nl}\right),
\end{aligned}$$

그리고

$$Var(S_x(D_{l,}YK_h)) = \frac{f(x)\nu_0}{nl}(\sigma^2(x) + m^2(x)) + o\left(\frac{1}{nl}\right).$$

□

4.2. 평활계수 선택과 SSNWR의 점근 성질

이 절에서는 AMISE의 수렴율을 최대화하는 평활계수와 그 때의 AMISE의 수렴율을 알아본다. 그리고 평활계수와 AMISE의 수렴율이 l과 u에 대한 조건에 따라 달라지는 내용도 알아본다. 연구 내용의 원활한 전개를 위해 다음과 같은 표기를 사용한다.

$$h_0 \sim l^{-a},\ h_l \sim l^{-b},\ h_u \sim u^{-c},\ h_t \sim n^{-d},\ a,b,c,d>0,$$

위의 정리에서 볼 수 있듯이 편의와 분산은 상충 (trade off)작용을 한다. 그러므로 AMISE를 최소화하는 평활계수는 두 값이 균형을 이룰 수 있도록 하는 값이다. 이러한 과정을 거치면 다음의 결과 얻을 수 있다.

Remark 4.1 $l \sim u \sim n$이라면, $\widehat{m}_{NW}$, $\widehat{m}_{ss1}$그리고 $\widehat{m}_{ss2}$의 AMISE 수렴율을 최대화하는 평활계수의 조건과 그 때의 AMISE는 다음과 같다.

a) $b=\frac{1}{5}$일 때 $AMISE(\widehat{m}_{NW}) \sim l^{-4/5}$.

b) $a \geq b$, $d=\frac{1}{5}$일 때 $AMISE(\widehat{m}_{ss1}) \sim l^{-4/5}$.

c) $b=c=\frac{1}{5}, a \geq c$일 때 $AMISE(\widehat{m}_{ss2}) \sim l^{-4/5}$.

표본크기 u와 l이 동일한 차수를 가진다면 ($l \sim u$), $\widehat{m}_{NW}$, $\widehat{m}_{ss1}$그리고 $\widehat{m}_{ss2}$는 모두 동일하게 $AMISE \sim n^{-4/5}$이 된다. 이 사실로부터 우리는 언라벨자료가 AMISE 수렴율을 개선하는데 도움이 되지 않는다는 것을 알 수 있다. 그리고 $a \geq b$와 $a \geq c$로부터 $h_0 = o(l^{-1/5})$이 되어야 한다는 것도 알 수 있다.

Remark 4.2 $u \sim l^p, p>1$이라면, $\widehat{m}_{NW}$, $\widehat{m}_{ss1}$그리고 $\widehat{m}_{ss2}$의 AMISE 수렴율을 최대화하는 평활계수의 조건과 그 때의 AMISE는 다음과 같다.

a) $b=\frac{1}{5}$일 때 $AMISE(\widehat{m}_{NW}) \sim l^{-4/5}$

b) b) $a \geq dp$, $d=\frac{1}{5}$일 때 $AMISE(\widehat{m}_{ss1}) \sim l^{-4p/5}$

c) $a \geq cp$, $1-b \geq p(1-c)$, $b=\frac{1}{5}$일 때 $AMISE(\widehat{m}_{ss2}) \sim l^{-4/5}$.
$a \geq cp$, $1-3b \leq p(1-3c)$,$b=1-\frac{4}{5}p$, $c=\frac{1}{5}$일 때 $AMISE(\widehat{m}_{ss2}) \sim l^{-4p/5}$.
$a \geq cp$, $1-b \leq p(1-c), 1-3b \geq p(1-3c)$, $b=\frac{1}{5}, c=1-\frac{4}{5p}$일 때 $AMISE(\widehat{m}_{ss2}) \sim l^{-4p/5}$.

여기에서 우리는 언라벨자료의 차수가 더 큰 경우, $u \sim l^p, p>1$에는 언라벨자료가 $\widehat{m}_{ss1}$와 $\widehat{m}_{ss2}$의 AMISE 수렴율을 빠르게 하는데 도움이 된다는 사실을 알 수 있다. $\widehat{m}_{ss1}$는 $a \geq dp$, $d=\frac{1}{5}$일 때, 그리고 $\widehat{m}_{ss2}$는 p에 따라서 적절한 b와 c를 선택함으로써 $AMISE \sim l^{-4p/5} (\sim u^{-4/5})$가 되게 할 수 있다. 그러므로 언라벨자료가 AMISE 수렴율을 빠르게 하는데 도움이 될 수 있다는 것을 알 수 있다. 그리고 초기추정에 사용되는 평활계수는 $l \sim u \sim n$일 때와 마찬가지로 $h_0 = o(l^{-1/5})$이 되어야 한다는 것도 알 수 있다 ($a \geq cp$, $a \geq dp$). 비록 대표본에서는 언라벨자료가 수렴율 개선에 도움이 되는 것으로 나타나지만 이를 적용하기에는 모수추정 등 해결해야 할 문제가 많아 실제 자료에 언라벨자료가 도움을 줄 수 있는 방법을 개발하는 연구가 필요하다.

$\widehat{m}_{ss2}$는 라벨자료와 언라벨자료의 주변분포를 다르게 가정한다. 그렇지만 이러한 가정은 추정량의 근사적인 성질에 직접적으로 영향을 미치지 않는 것으로 나타난다.

5. 결론

본 논문에서는 새로운 준지도 회귀모형을 제안하고 평활계수 선택의 기준을 제시하였다. 제안된 모형은 좀 더 일반화된 방법으로 라벨자료와 언라벨자료에 사용되는 평활계수를 서로 다른 값으로 간주하였다. 제안된 방법의 점근분포를 규명하고, 이를 바탕으로 AMISE의 수렴율을 최대로 할 수 있는 평활계수 조건에 대해 알아보았는데, 적절한 평활계수가 선택된다면 준지도 회귀모형의 수렴율이 더 좋다는 것을 입증하였다. 그리고 언라벨자료에 대한 초기추정은 과적합되는 것이 더 좋음을 증명하였다.

Nadaraya-Watson 추정량을 본 연구에서 사용한 것은 이론적 증명의 편의를 위함이었다. 그렇지만 본 연구를 바탕으로 국소다항회귀 (local polynomial regression model), 가법모형 (additive model) 혹은 부스팅나무 (boosting tree)에 기반한 준지도 회귀모형을 개발하고 더 정확한 모수선택에 대해 알아보는 것도 흥미있는 연구주제가 될 수 있을 것이다.

본 연구에서 입증한 내용을 소표본 대한 모의실험을 통해 확인하는 과정을 거치는 것이 좋을 것으로 생각되었다. 그렇지만, 모수 선택에 필요한 여러가지 값과 자료의 크기 등이 추정에 많은 영향을 끼치게 되어 정확한 모의실험을 설계하는 것이 쉽지 않아 추후 연구 과제로 남긴다.

References

Belkin, M., Niyogi, P. and Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *The Journal of Machine Learning Research*, **7**, 2399-2434.

Chapelle, O., Scholkopf, B. and Zien, A. (2006). *Semi-supervised learning*, MIT Press, Cambridge, MA.

Cortes, C. and Mohri, M. (2007). On transductive regression. *Advances in Neural Information Processing System*, **19**, 305-312.

Liu, B., Jing, L., Yu, J. and Jia L. (2014). Constrained least squares regression for semi-supervised learning. In *Advances in Knowledge Discovery and Data Mining*, **8444**, 110-121.

Lafferty, J. and Wasserman, L. (2008). Statistical analysis of semi-supervised regression. In *Advances in Neural Information Processing Systems*, **20**, 801-808.

Nadaraya, E. A. (1964). *On estimating regression. Theory of Probability and its Applications*, **9**, 141-142.

Niyogi, P. (2008). *Manifold regularization and semi-supervised learning: Some theoretical analyses*, Technical Report TR-2008-01, Computer science department, University of Chicago, Chicago, IL.

Seok, K. (2012). Study on semi-supervised local constant regression estimation. *Journal of the Korean Data & Information Science Society*, **23**, 579-585.

Seok, K. (2013). A study on semi-supervised kernel ridge regression estimation. *Journal of the Korean Data & Information Science Society*, **24**, 341-353.

Seok, K. (2015). Semisupervised support vector quantile regression. *Journal of the Korean Data & Information Science Society*, **26**, 517-524.

Suykens, J.A.K., Gastel, T. V., Bravanter, J. D., Moore, B. D. and Vandewalle, J. (2002). *Least squares support vector machines*, World Scientific, London.

Wang, M., Hua, X., Song, Y., Dai, L. and Zhang, H. (2006). Semi-supervised kernel regression. In *Proceeding of the Sixth International Conference on Data Mining*, 1130-1135.

Wasserman, L. (2006). *All of nonparametric statistics*, Springer, New York.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of StatisticsA*, **26**, 359-372.

Wei, R., Pan, L. and Guo, L. (2015). Semi-supervised learning via nonnegative least squares regression. In *Proceedings of the 7th International Conference on Internet Multimedia Computing and Service*, **15**, 105-116.

Xu, S., An. X., Qiao, X., Zhu, L. and Li, L. (2011). Semisupervised least squares support vector regression machines. *Journal of Information & Computational Science*, **8**, 885-892.

Xu, Z., King, I. and Lyu, M. R. (2010). *More than semi-supervised learning*, LAP LAMBERT Academic Publishing, London.

Zhu, D. (2005). *Semi-supervised learning literature survey*, Technical Report, Computer Sciences Department, University of Wisconsin, Madison, WI.

Zhu, X. and Goldberg, A. (2009). *Introduction to semi-supervised learning*, Morgan & Claypool, London.

Journal of the Korean Data & Information Science Society
2016, **27**(4), 993–1000

http://dx.doi.org/10.7465/jkdi.2016.27.4.993
한국데이터정보과학회지

Smoothing parameter selection in semi-supervised learning[†]

Kyungha Seok[1]

[1]Department of Statistics, Inje University

Received 16 May 2016, revised 16 June 2016, accepted 22 July 2016

Abstract

Semi-supervised learning makes it easy to use an unlabeled data in the supervised learning such as classification. Applying the semi-supervised learning on the regression analysis, we propose two methods for a better regression function estimation. The proposed methods have been assumed different marginal densities of independent variables and different smoothing parameters in unlabeled and labeled data. We shows that the overfitted pilot estimator should be used to achieve the fastest convergence rate and unlabeled data may help to improve the convergence rate with well estimated smoothing parameters. We also find the conditions of smoothing parameters to achieve optimal convergence rate.

Keywords: Asymptotic mean integrated squared error, convergence rate, kernel regression, Nadaraya-Watson estimator, semi-supervised regression, smoothing parameter.

[†] 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 (2011-0009705).
[1] Professor, Department of Statistics and Institute of Statistical Information, Inje University, Kimhae, 50834, Korea. E-mail: statskh@inje.ac.kr