The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.1.131
26(1), 131–139

Comparison of Feature Selection Methods in Support Vector Machines

Kwangsu Kim[a] · Changyi Park[a,1]

[a]Department of Statistics, University of Seoul

(Received December 15, 2012; Revised January 16, 2013; Accepted January 16, 2013)

Abstract

Support vector machines(SVM) may perform poorly in the presence of noise variables; in addition, it is difficult to identify the importance of each variable in the resulting classifier. A feature selection can improve the interpretability and the accuracy of SVM. Most existing studies concern feature selection in the linear SVM through penalty functions yielding sparse solutions. Note that one usually adopts nonlinear kernels for the accuracy of classification in practice. Hence feature selection is still desirable for nonlinear SVMs. In this paper, we compare the performances of nonlinear feature selection methods such as component selection and smoothing operator(COSSO) and kernel iterative feature extraction(KNIFE) on simulated and real data sets.

Keywords: Component selection and smoothing operator, kernel iterative feature extraction, support vector machines.

1. 서론

Cortes와 Vapnik (1995)가 제안한 지지벡터기계(support vector machines; SVM)는 주어진 자료를 가장 잘 분류하기 위하여 마진(margin)이 가장 큰 초평면을 찾아 분류하는 학습법으로서 예측력이 매우 높고 모형의 변형이 쉬우며 고차원 자료(high dimensional data)를 잘 다룰수 있기 때문에 여러 가지 분류문제에서 성공적으로 적용되고 있다. 그러나 일반적인 지지벡터기계는 잡음변수(noise variable)가 많은 경우에 오분류율이 증가할 수 있고 최종 분류기(classifier)에서 각 변수의 중요도를 파악하기가 어렵다. 이러한 문제점을 보완하기 위하여 잡음변수는 제거하고 설명력이 높은 변수들만으로 최종 분류기를 구성하도록 하는 변수선택법을 고려할 수 있다.

선형 지지벡터기계에서는 회귀분석의 변수선택법을 자연스럽게 확장할 수 있다. 예를 들면 부분집합선택법의 일종인 Guyon 등 (2002)의 SVM-RFE(recursive feature elimination), 벌점화를 이용한 방법인 LASSO(least absolute shrinkage and selection operator)를 적용한 L1-SVM (Zhu 등, 2003)과

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (No. 2012R1A1A2004901).

[1]Corresponding author: Associate Professor, Department of Statistics, University of Seoul, 90 Jeonnong-Dong, Dongdaemun-gu, Seoul 130-743, Korea. E-mail: park463@uos.ac.kr

SCAD(smoothly clipped absolute deviation)를 적용한 SCAD-SVM (Zhang 등, 2006) 등 선형 지지벡터기계의 변수선택법에 대한 여러 가지 연구가 있다. 일반적으로 분류의 정확도를 높이기 위해서 선형 대신 비선형 커널을 사용하는 것이 관례이며, 비선형 분류 문제에서도 역시 예측력의 손실없이 최종 분류기에서 사용되는 입력변수의 갯수를 줄일 필요가 있다.

선형의 경우에는 계수값에 적절한 벌점을 적용하면 입력변수들을 선택할 수 있지만 비선형에서는 이러한 방법을 직접 적용하면 입력변수가 아닌 커널에 의한 기저함수가 선택되는 난점이 있다. 즉 비선형 커널의 경우에는 고차원의 특성공간(feature space)의 차원 축소가 목표가 아니라 입력공간의 축소가 목표가 된다. 최근 비선형 지지벡터기계의 변수선택법으로 Zhang (2006)이 제안한 COSSO(component selection and smoothing operator)와 Allen (2011)이 제안한 KNIFE(kernel iterative feature extraction)가 대표적이다. 본 논문에서는 비선형 지지벡터기계의 변수선택법인 COSSO와 KNIFE를 여러 가지 모의실험 및 실제 자료를 통하여 비교하고자 한다.

본 논문은 다음과 같이 구성되어 있다. 2절에서는 지지벡터기계와 비선형 변수선택법인 COSSO와 KNIFE에 대하여 설명한다. 3절에서는 모의실험과 실제자료를 통하여 변수선택 및 예측력을 비교하며, 마지막으로 4절에서는 결론을 기술한다.

2. 비선형 지지벡터기계의 변수선택법

2.1. 지지벡터기계

본 논문에서는 입력공간이 $\mathcal{X} = \prod_{j=1}^{p} \mathcal{X}_j \subset \mathbb{R}^p$, 출력공간이 $\mathcal{Y} = \{-1, +1\}$인 이진 분류문제만을 고려하기로 한다. 고려 대상이 되는 함수공간은 $\mathcal{F}$, 훈련자료는 $(\boldsymbol{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}, i = 1, \ldots, n$이고 훈련자료를 이용하여 추정된 의사결정함수를 $\hat{f} \in \mathcal{F}$라 하자. 그러면 분류기 $\text{sign}(\hat{f}(\boldsymbol{x}))$를 이용하여 $\boldsymbol{x}$에서 y의 값을 예측할 수 있다.

지지벡터기계를 설명하기 위해서 몇 가지 기호를 더 도입하기로 한다. $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$은 선형 또는 비선형의 커널이고 $\mathcal{H}_K$는 커널 K에 의해 생성되는 RKHS(reproducing kernel Hilbert space)를 나타낸다. 이 경우 지지벡터기계에서 고려하는 의사결정함수는 $f(\boldsymbol{x}) = h(\boldsymbol{x}) + b$의 형태이다. 단 $h \in \mathcal{H}_K$이고 $b \in \mathbb{R}$이다.

일반적으로 지지벡터기계는 다음과 같은 목적함수를 최소화하는 $f \in \mathcal{F}$를 찾는다.

$$\frac{1}{n} \sum_{i=1}^{n} [1 - y_i f(\boldsymbol{x}_i)]_+ + \frac{\lambda}{2} \|h\|^2. \tag{2.1}$$

단, $[z]_+ = \max(z, 0)$, $\lambda > 0$는 벌점모수, $\|\cdot\|$는 커널 K에 의해 정의되는 커널 노음(kernel norm)이다. Representer 정리에 따르면 식 (2.1)의 해는 적절한 $\hat{\alpha}$와 $\hat{b}$에 대하여

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} \hat{\alpha}_i K(\boldsymbol{x}, \boldsymbol{x}_i) + \hat{b}$$

로 나타낼 수 있다 (Wahba, 1990). 모수의 추정 등 자세한 사항은 Cortes와 Vapnik (1995)를 참고하기 바란다.

2.2. COSSO

Zhang (2006)이 제안한 COSSO는 지지벡터기계의 의사결정함수 $f \in \mathcal{F}$에 대하여 다음과 같은 함수

ANOVA 분해(functional analysis of variance decomposition)를 적용한다.

$$f(\boldsymbol{x}) = b + \sum_{j=1}^{p} f_j(x_j) + \sum_{j<k} f_{jk}(x_j, x_k) + \cdots, \tag{2.2}$$

여기서 b는 상수항, f_j는 변수 x_j의 주효과, f_{jk}는 변수 x_j와 x_k의 이인자 교호작용에 해당한다. 식 (2.2)에 대응되는 함수공간 $\mathcal{F}$의 ANOVA 분해는

$$\mathcal{H} = \{1\} \oplus \sum_{j=1}^{p} \bar{\mathcal{H}}_j \oplus \sum_{j<k} \left(\bar{\mathcal{H}}_j \otimes \bar{\mathcal{H}}_k\right) \oplus \cdots \tag{2.3}$$

로 표현된다. 여기서 $\bar{\mathcal{H}}_j$는 상수공간 $\{1\}$에 직교하는 $\mathcal{X}_j$상의 RKHS의 부분공간으로 정의된다.

흔히 특정 차수 이하의 교호작용만을 고려하기 때문에 식 (2.3)에서 고려되지 않는 고차의 교호작용에 대응되는 공간을 제외한 부분공간을 $\mathcal{F}_\nu$, $\nu = 1, \ldots, d$로 나타내고

$$\mathcal{F} = \{1\} \oplus_{\nu=1}^{d} \mathcal{F}_\nu$$

로 재정의 한다. 그러면 재정의된 $\mathcal{F}$상의 (2.1)은 다음과 같다.

$$\begin{aligned} &\text{minimize } \frac{1}{n}\sum_{i=1}^{n}[1 - y_i f(\boldsymbol{x}_i)]_+ + \lambda \sum_{\nu=1}^{d} \theta_\nu^{-1} ||P^\nu f||^2 \\ &\text{subject to} \quad \lambda > 0,\ \theta_\nu \geq 0,\ \nu = 1, \ldots, d. \end{aligned} \tag{2.4}$$

단 $\lambda > 0$는 벌점모수, $\|\cdot\|$은 $\mathcal{F}$상의 커널 노음, P^ν는 $\mathcal{F}_\nu$상의 사영연산자, θ_ν는 스케일 모수이다.

COSSO-SVM은 다음과 같이 스케일 모수 θ_ν에 LASSO 타입의 벌점을 적용함으로써 설명력 있는 함수성분(functional component)을 선택한다.

$$\begin{aligned} &\text{minimize } \frac{1}{n}\sum_{i=1}^{n}[1 - y_i f(\boldsymbol{x}_i)]_+ + \lambda \sum_{\nu=1}^{d} \theta_\nu^{-1} ||h_\nu||^2 + \lambda_\theta \sum_{\nu=1}^{d} \theta_\nu \\ &\text{subject to} \quad \lambda > 0,\ \lambda_\theta > 0,\ \theta_\nu \geq 0,\ \nu = 1, \ldots, d. \end{aligned} \tag{2.5}$$

식 (2.5)의 해는 $f(\boldsymbol{x}) = b + \sum_\nu h_\nu(\boldsymbol{x})$ 형태로 $h_\nu(\boldsymbol{x}) = \theta_\nu \sum_{i=1}^{n} c_i K_\nu(\boldsymbol{x}_i, \boldsymbol{x})$는 ν번째 성분을 나타낸다. 단 K_ν는 F_ν상의 커널을 나타낸다. 쌍대 목적함수는 볼록이 아니며 θ_ν들에 대한 선형계획(linear program)과 b, c_i들에 대한 이차계획(quadratic program)으로 이루어진 이중볼록(bi-convex) 함수형태이므로 적절한 초기치로부터 출발하여 선형계획과 이차계획의 반복으로 해를 구하게 된다. 해를 구하는 방법에 대한 자세한 사항은 Zhang (2006)을 참조하기 바란다.

2.3. KNIFE

Allen (2011)이 제안한 KNIFE의 아이디어는 커널에서 변수에 대하여 가중치를 적용하고 가중치에 LASSO 벌점을 적용하여 설명력 높은 변수가 자동적으로 선택되도록 하는 것이다. 즉, $\boldsymbol{w}$를 영 또는 양수인 p차원 가중치 벡터라고 하면, 다음과 같은 가중 커널을 고려하게 된다.

- 선형커널: $K_{\boldsymbol{w}}(\boldsymbol{u}, \boldsymbol{v}) = \sum_{j=1}^{p}(w_j u_j)(w_j v_j)$.
- 다항커널: $K_{\boldsymbol{w}}(\boldsymbol{u}, \boldsymbol{v}) = (\sum_{j=1}^{p}(w_j u_j)(w_j v_j) + c)^d$. 단, $c \in \mathbb{R}$이고 d는 자연수이다.

• 가우스커널: $K_{\boldsymbol{w}}(\boldsymbol{u},\boldsymbol{v}) = \exp(-\gamma\sum_{j=1}^{p}(w_ju_j - w_jv_j)^2)$. 단, $\gamma > 0$이다.

$\boldsymbol{K_w} = (K_{\boldsymbol{w}}(\boldsymbol{x}_i,\boldsymbol{x}_{i'})) \in \mathbb{R}^{n\times n}$은 가중치 행렬이고 대응되는 의사결정 함수는 $f_{\boldsymbol{w}}(\boldsymbol{x}) = \sum_{i=1}^{n} c_iK_{\boldsymbol{w}}(\boldsymbol{x}_i,\boldsymbol{x})$로 나타내자. 지지벡터기계의 목적함수 (2.1)에서 가중 커널과 가중치에 대한 LASSO 벌점항을 추가하면 다음과 같은 KNIFE 목적함수를 얻는다.

$$\begin{aligned}&\text{minimize}_{\boldsymbol{c},\boldsymbol{w}}\frac{1}{n}\sum_{i=1}^{n}[1-y_if_{\boldsymbol{w}}(\boldsymbol{x}_i)]_+ + \lambda\boldsymbol{c}^T\boldsymbol{K_w}\boldsymbol{c} + \lambda_{\boldsymbol{w}}\mathbf{1}^T\boldsymbol{w} \\ &\text{subject to}\quad \lambda > 0,\ \lambda_{\boldsymbol{w}} \geq 0,\ 0 \leq w_j < 1,\ j = 1,\ldots,p.\end{aligned} \tag{2.6}$$

선형커널의 경우 $\mathbf{K_w} = \mathbf{XWX}^T$으로 정의하고 $\mathbf{W} = \text{diag}(\mathbf{w}^2)$이고 $\boldsymbol{\beta} = \mathbf{X}^T\boldsymbol{c}$라고 하자. 그러면 KNIFE의 목적함수는 다음과 같이 쓸 수 있다.

$$\begin{aligned}&\text{minimize}_{\boldsymbol{\beta},\boldsymbol{w}}\frac{1}{n}\sum_{i=1}^{n}[1-y_if_{\boldsymbol{w}}(\boldsymbol{x}_i)]_+ + \lambda\boldsymbol{\beta}^T\boldsymbol{W}\boldsymbol{c} + \lambda_{\boldsymbol{w}}\mathbf{1}^T\boldsymbol{w} \\ &\text{subject to } \lambda > 0, \lambda_{\boldsymbol{w}} \geq 0, 0 \leq w_j < 1, j = 1,\ldots,p.\end{aligned} \tag{2.7}$$

이 목적함수 (2.7)은 COSSO와 마찬가지로 볼록함수가 아닌 이중볼록함수이므로 적절한 초기치로부터 $\boldsymbol{c}$와 $\boldsymbol{w}$를 번갈아가며 고정시킨후에 최적화하게 된다. 비선형커널인 경우에는 목적함수의 손실함수 부분이 $\boldsymbol{w}$의 볼록함수가 아니므로 커널을 다음과 같이 선형 근사한다.

$$\tilde{K}_{\boldsymbol{w}^{(t)}}(\boldsymbol{x}_i,\boldsymbol{x}_j) = K_{\boldsymbol{w}}^{(t-1)}(\boldsymbol{x}_i,\boldsymbol{x}_j) + \triangledown K_{\boldsymbol{w}^{(t-1)}}(\boldsymbol{x}_i,\boldsymbol{x}_j)^T\left(\boldsymbol{w}^{(t)} - \boldsymbol{w}^{(t-1)}\right).$$

단, $\boldsymbol{w}^{(t-1)}$은 $t-1$단계에서 추정된 가중치 벡터를 나타낸다. $\boldsymbol{A} = (A_{ij}) \in \mathbb{R}^{n\times n} : A_{ij} = \sum_{j=1}^{n} c_j \triangledown K_{\boldsymbol{w}^{(t-1)}}(\boldsymbol{x}_i,\boldsymbol{x}_j)^T$와 $\boldsymbol{B} = (B_{ij}) \in \mathbb{R}^{n\times n} : B_{ij} = K_{\boldsymbol{w}^{(t-1)}}(\boldsymbol{x}_i,\boldsymbol{x}_j) - \triangledown K_{\boldsymbol{w}^{(t-1)}}^T\boldsymbol{w}^{(t-1)}$이라 하자. 식 (2.6)에서 K대신 선형근사 $\tilde{K}$를 적용하면

$$\begin{aligned}&\text{minimize}_{\boldsymbol{c},\boldsymbol{w}}\frac{1}{n}\sum_{i=1}^{n}[1-y_i(\boldsymbol{B}\boldsymbol{c}+\boldsymbol{A}\boldsymbol{w})_i]_+ + \lambda\boldsymbol{c}^T\boldsymbol{A}\boldsymbol{w} + \lambda_{\boldsymbol{w}}\mathbf{1}^T\boldsymbol{w} \\ &\text{subject to}\quad \lambda > 0,\ \lambda_{\boldsymbol{w}} \geq 0,\ 0 \leq w_j < 1,\ j = 1,\ldots,p\end{aligned} \tag{2.8}$$

이며 $\boldsymbol{c}$와 $\boldsymbol{w}$의 이중볼록함수가 되며 선형커널의 경우와 마찬가지로 $\boldsymbol{c}$와 $\boldsymbol{w}$에 대하여 번갈아가며 반복적으로 최적화를 한다. 해를 구하는 방법에 대한 자세한 사항은 Allen (2011)을 참조하기 바란다.

3. 자료 분석

이진분류에 대한 COSSO는 R 패키지 `COSSO`에 구현되어 있으나 스플라인 커널만을 사용하도록 되어 있는 반면, COSSO의 다범주로의 확장인 Lee 등 (2006)의 `SMSVM`에서는 여러 가지 커널을 사용할 수 있기 때문에 본 논문의 자료 분석에 사용하였다. 참고로 `SMSVM`의 R 소스 코드는 `http://www.stat.osu.edu/~yklee/software.html`에서 내려 받을 수 있다. KNIFE의 경우에는 `http://www.stat.rice.edu/~gallen/software.html`의 Matlab 소스 코드를 이용하여 자료를 분석하였다.

3.1. 모의 실험

COSSO와 KNIFE의 예측력과 변수선택을 비교하기 위해 다음과 같은 두 가지 모의실험을 고려하였다.

Table 3.1. Results for simulation A: prediction

p	n	Test Error (S.E.)	
		COSSO	KNIFE
4	50	0.2453 (0.0003)	0.2445 (0.0028)
	100	0.2231 (0.0001)	0.2192 (0.0017)
	200	0.2096 (0.0001)	0.2057 (0.0007)
10	50	0.2719 (0.0007)	0.2516 (0.0031)
	100	0.2333 (0.0002)	0.2316 (0.0016)
	200	0.2116 (0.0001)	0.2245 (0.0007)

Table 3.2. Results for simulation A: feature selection

p	n	COSSO			KNIFE		
		C	IC	PPS	C	IC	PPS
4	50	1.84	1.16	0.17	1.74	0.60	0.37
	100	1.99	1.29	0.15	1.86	0.71	0.41
	200	2.00	1.48	0.08	1.99	0.84	0.43
10	50	1.77	3.26	0.13	1.63	1.49	0.24
	100	1.95	3.70	0.10	1.93	2.55	0.25
	200	2.00	4.07	0.05	2.00	2.81	0.28

Note: C and IC denote the number of correctly and incorrectly selected input variables, respectively, and PPS is the probability of selecting the true model.

- 모의실험 A

  입력변수 $X_1, \ldots, X_p$는 서로 독립적으로 $N(0,1)$을 따르며, 입력변수 $\boldsymbol{X} = (X_1, \ldots, X_p)^T$의 값이 $\boldsymbol{x}$로 주어졌을때 $Y = 1$의 조건부확률에 대하여 다음과 같은 로지스틱 모형을 가정한다.

$$\log\left(\frac{\mathbb{P}(Y=+1|\boldsymbol{X}=\boldsymbol{x})}{\mathbb{P}(Y=-1|\boldsymbol{X}=\boldsymbol{x})}\right) = x_1 + 2x_2.$$

  이 경우 X_1과 X_2는 설명력이 있는 변수이고 나머지 $X_3, \ldots, X_p$는 잡음변수가 된다. 모의실험 A에서는 COSSO와 KNIFE에서 선형커널을 적용하기로 한다.

- 모의실험 B

  모의실험 A와 마찬가지로 입력변수 $X_1, \ldots, X_p$는 서로 독립적으로 $N(0,1)$을 따르며, 입력변수가 $\boldsymbol{X} = \boldsymbol{x}$로 주어졌을때 $Y = 1$의 조건부확률에 로지스틱 모형

$$\log\left(\frac{\mathbb{P}(Y=+1|\boldsymbol{X}=\boldsymbol{x})}{\mathbb{P}(Y=-1|\boldsymbol{X}=\boldsymbol{x})}\right) = x_1 + \pi\sin(\pi\, x_2)$$

  을 가정한다. 모의실험 A와 같이 설명력이 있는 변수는 X_1과 X_2이고 나머지는 잡음변수이며 COSSO와 KNIFE에서 가우스 커널을 적용하기로 한다.

각 모의실험에서 $p\,(= 4, 10)$개의 입력변수를 생성하고 $n\,(= 50, 100, 200)$개의 훈련자료와 10,000개의 시험자료를 생성하였다. 조율모수(tuning parameter)의 값을 결정하기 위하여 검증오차에 대한 5-묶음 교차확인법(5-fold crossvalidation)을 적용하였다. 또한 실험의 변동성을 파악하기 위하여 이러한 과정을 100회 반복 실시하였다.

Table 3.1–3.4는 각각 모의실험 A와 모의실험 B에서 100회 반복 시행에 대한 COSSO와 KNIFE의 시험자료의 예측력(평균 오분류율과 표준오차) 및 변수선택(C, IC, PPS)을 비교한 결과이다. 두 모의실

Table 3.3. Results for simulation B: prediction

p	n	Test Error (S.E.)	
		COSSO	KNIFE
4	50	0.3422 (0.0009)	0.3274 (0.0089)
	100	0.2373 (0.0005)	0.2342 (0.0035)
	200	0.1934 (0.0001)	0.1954 (0.0013)
10	50	0.4490 (0.0006)	0.4299 (0.0093)
	100	0.3545 (0.0010)	0.3266 (0.0113)
	200	0.2127 (0.0003)	0.1945 (0.0027)

Table 3.4. Results for simulation B: feature selection

p	n	COSSO			KNIFE		
		C	IC	PPS	C	IC	PPS
4	50	1.68	1.28	0.08	1.31	0.34	0.18
	100	1.97	1.69	0.04	1.51	0.32	0.30
	200	2.00	1.89	0.02	1.91	0.30	0.69
10	50	1.10	3.24	0.04	1.05	2.70	0.03
	100	1.68	4.21	0.03	1.04	0.66	0.05
	200	2.00	6.25	0.02	1.65	0.58	0.39

험에서 공통적으로 고정된 p에 대하여 n이 증가하면 설명력 있는 변수의 갯수 C는 증가하는 경향이 있고, 고정된 n에 대하여 p가 증가하면 전체적으로 IC나 오분류율이 증가하는 경향이 있다. COSSO는 각 p에 대하여 n이 증가함에 따라 반드시 IC는 감소하고 PPS가 증가하지는 않으며 반면 KNIFE는 IC는 감소하고 PPS는 증가하는 경향을 보인다. 전반적으로 COSSO는 KNIFE에 비해 잡음변수를 더 많이 선택하는 경향이 있고 예측력 측면에서도 잡음변수의 선택에 의하여 오분류율이 증가하는 것으로 보인다.

3.2. 실제 자료

COSSO와 KNIFE의 비교를 위해 다음과 같은 UCI Machine Learning Repository 자료를 분석하였다.

- BUPA 자료

  자료의 갯수는 345이고 입력변수의 갯수는 6으로 과도한 음주로 인한 간질환에 대하여 민감한 혈액검사 결과인 mcv(mean corpuscular volume), alkphos(alkaline phosphotase), sgpt(alamine aminotransferase), sgot(aspartate aminotransferase) gammagt(gamma-glutamyl transpeptidase)와 하루에 반 파인트(pint) 단위로 몇 잔을 마시는 지를 나타내는 drinks로 이루어져 있다.

- Ionosphere 자료

  자료의 수는 351개이고 입력변수는 34개로 이루어져 있다. 출력변수의 값은 전리층에서 레이더의 반환값에 어떤 구조가 있는 경우는 Good의 값을 갖고 아무런 구조가 없는 경우에는 신호가 그대로 전리층을 통과하며 Bad의 값을 갖는다.

- Sonar 자료

  자료의 수는 208개이고 입력변수의 갯수는 60으로 각 입력변수는 0과 1사이의 값을 가지며 여러 가지 각도나 조건하에서의 광물(M) 또는 암석(R)의 소나(sonar) 반사파 신호 패턴을 나타낸다. 출력변수는 물체가 암석인 경우 R의 값을 갖고 광물인 경우에는 M의 값을 갖는다.

Table 3.5. Predictive performances of COSSO and KNIFE on UCI benchmark data sets

Kernel	Method	BUPA	Ionosphere	Sonar
Linear	COSSO	0.3181 (0.0053)	0.1381 (0.0039)	0.2420 (0.0076)
	KNIFE	0.3186 (0.0052)	0.1385 (0.0036)	0.2710 (0.0075)
Gaussian	COSSO	0.3702 (0.0061)	0.0796 (0.0035)	0.1982 (0.0075)
	KNIFE	0.3683 (0.0082)	0.0697 (0.0035)	0.2501 (0.0086)

Table 3.6. Feature selection results of COSSO and KNIFE on UCI benchmark data sets

Kernel	Method	BUPA	Ionosphere	Sonar
Linear	COSSO	5.80 (0.057)	19.86 (1.235)	39.94 (1.921)
	KNIFE	5.94 (0.044)	14.20 (1.131)	18.86 (1.719)
Gaussian	COSSO	5.84 (0.104)	24.40 (1.251)	40.24 (1.765)
	KNIFE	3.82 (0.223)	11.38 (0.414)	19.94 (1.731)

COSSO와 KNIFE의 예측력과 변수선택을 비교하기 위하여 전체 자료를 랜덤하게 2:1의 비율로 훈련자료와 시험자료로 분할하여 훈련자료로는 모형의 적합 및 최적화를 하고 시험자료에 대한 오분류율을 구하였다. 모형의 최적화 과정에서는 모의실험과 마찬가지로 5-묶음 교차확인법을 적용하였다. 또한 변동성을 파악하기 위하여 랜덤 분할을 50회 반복하여 실험하였다.

Table 3.5과 Table 3.6는 각각 세 가지 UCI 자료에 대하여 50회의 랜덤 분할에 대한 예측력과 변수선택 결과를 비교한다. BUPA와 Ionosphere 자료의 경우 예측력면에서는 COSSO나 KNIFE는 큰 차이를 보이지 않는다. 다만 KNIFE의 경우 COSSO보다 평균적으로 적은 수의 변수를 선택하므로 보다 효율적으로 변수를 선택한다고 할 수 있다. Sonar 자료의 경우에 KNIFE는 COSSO보다 절반정도의 변수를 선택하는 반면 오분류율이 더 높게 나타났다. 이 자료의 경우에는 대부분의 입력변수들이 설명력이 있기 때문에 전반적으로 적은 수의 변수를 선택하는 경향이 있는 KNIFE의 예측성능이 떨어지는 것으로 추측할 수 있다. 커널에 따른 예측력과 선택된 변수 갯수는 자료 특성에 의존하는 것으로 보인다.

4. 결론

본 논문에서는 비선형 지지벡터기계에서 변수선택법에 대하여 모의실험과 실제 자료분석을 통하여 COSSO와 KNIFE의 예측력과 변수선택에 대하여 비교하였다. COSSO의 경우 RHKS에 대한 함수 ANOVA 분해와 스케일 모수에 대한 LASSO 벌점화를 통한 비선형 성분의 선택을 고려하며 KNIFE는 가중 커널과 가중치에 대한 LASSO 벌점화를 이용한다. 두 방법 모두 특성공간이 아닌 원래의 입력공간상에서 차원 축소를 한다는 점에서 비선형 지지벡터기계에 대한 적절한 변수 선택법으로 볼 수 있다. 개념상으로는 RHKS 함수공간의 분해를 고려하는 COSSO가 더 자연스러워 보인다. 하지만 모의실험과 실제 자료분석의 결과에 따르면 KNIFE가 COSSO보다 더 적은 수의 변수들을 선택함을 볼 수 있었다. 따라서 설명력있는 변수들의 갯수가 입력공간의 차원 p에 비해 적은 경우에는 KNIFE가 COSSO보다 더 좋은 예측과 변수선택을 할 것으로 예상된다. 또한 두 방법 모두 이중볼록함수를 번갈아가며 반복적으로 최적화하는 방식으로 해를 구하는데 변수선택 측면에서 일치성이 성립할 지는 이론적으로 연구해 볼 필요가 있다.

References

Allen, G. I. (2011). Automatic feature selection via weighted kernels and regularization, *Journal of Computational and Graphical Statistics*, In Press.

Cortes, C. and Vapnik, V. (1995). Support-vector networks, *Machine Learning*, **20**, 273–297.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines, *Machine Learning*, **46**, 389–422.

Lee, Y., Kim, Y., Lee, S. and Koo, J.-Y. (2006). Structured multicategory support vector machines with analysis of variance decomposition, *Biometrika*, **93**, 555–571.

Wahba, G. (1990). Spline methods for observational data, *CBMS-NSF Regional Conference Series*, Philadelphia.

Zhang, H. (2006). Variable selection for SVM via smoothing spline ANOVA, *Statistica Sinica*, **16**, 659–674.

Zhang, H., Ahn, J., Lin, X. and Park, C. (2006). Gene selection using support vector machines with nonconvex penalty, *Bioinformatics*, **22**, 185–202.

Zhu, J., Rosset, S., Hastie, T. and Tibshirani, R. (2003). 1-norm support vector machines, *Neural Information Processing Systems*, MIT Press.

지지벡터기계의 변수 선택방법 비교

김광수[a] · 박창이[a,1]

[a]서울시립대학교 통계학과

(2012년 12월 15일 접수, 2013년 1월 16일 수정, 2013년 1월 16일 채택)

요 약

지지벡터기계는 잡음변수가 존재하는 경우에 성능이 저하될 수 있다. 또한 최종 분류기에서 각 변수들의 중요도를 알리 어려운 단점이 있다. 따라서 변수선택은 지지벡터기계의 해석력과 정확도를 높일 수 있다. 기존의 문헌상의 대부분의 연구는 선형 지지벡터기계에서 성근 해를 주는 벌점함수를 통해 변수를 선택에 관한 것이다. 실제로는 분류의 정확도를 높이기 위해 비선형 커널을 사용하는 경우가 일반적이다. 따라서 변수선택은 비선형 지지벡터기계에서도 마찬가지로 필요하다. 본 논문에서는 모의실험 및 실제자료를 통하여 비선형 지지벡터의 대표적인 변수선택법인 COSSO(component selection and smoothing operator)와 KNIFE(kernel iterative feature extraction)의 성능을 비교한다.

주요용어: COSSO, KNIFE, 비선형 커널.

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2012R1A1A2004901).

[1]교신저자: (130-743) 서울시 동대문구 시립대길 163, 서울시립대학교 통계학과, 교수.
E-mail: park463@uos.ac.kr