The Korean Communications
in Statistics Vol. 10 No. 3, 2003
pp. 873-878

# A Study on Support Vectors of Least Squares Support Vector Machine[1)]

**Kyungha Seok[2)]  Daehyun Cho[3)]**

## Abstract

LS-SVM(Least-Squares Support Vector Machine) has been used as a promising method for regression as well as classification. Suykens et al.(2000) used only the magnitude of residuals to obtain SVs(Support Vectors). Suykens' method behaves well for homogeneous model. But in a heteroscedastic model, the method shows a poor behavior. The present paper proposes a new method to get SVs. The proposed method uses the variance of noise as well as the magnitude of residuals to obtain support vectors. Through the simulation study we justified excellence of our proposed method.

*Keywords* : Support Vector, Least-Squares Support Vector Machine, Heteroscedastic model.

## 1. 서론

최근 수십년 동안 신경망은 분류(classification)과 회귀분석(regression)등 많은 분야에서 각광을 받아 왔다(Bishop(1995), Cherkassky 등(1998), Haykin(1994)). 그럼에도 불구하고 은닉층과 은닉 노드의 수를 결정할 때 목적함수가 여러 개의 국소 최소점을 가지므로 최적의 해를 찾는 것이 큰 문젯거리로 대두되어 왔다.

최근에 신경망의 일종인 SVM(Support Vector Machines)이 개발되어 기존의 신경망이 가지는 단점을 극복하였다(Vapnik(1998), Cristianini 등(2000)). 이는 목적함수가 볼록함수형태(convex)이기 때문에 국소 최소에 신경을 쓰지 않아도 될 뿐 아니라 모형의 복잡도(complexity)도 SVM을 해결하는 과정에서 만들어지는 써포트 벡터(SV)로 해결이 된다.

많은 논문에서 여러가지 자료를 통해 SVM의 우월성을 증명하고 있다(Vapnik(1998), Cristianini 등(2000)). 그러나 SVM을 훈련시키기 위해서는 QP(Quadratic Programming)문제를 해결해야 하는 단점이 있다. 이는 계산에 많은 시간과 저장공간을 요구한다. 실제로 규모가 큰 자료에서는 SVM을 훈련하는데 2~3일 정도가 소요되는 것으로 알려져 있다. 이러한 단점 때문에 SVM이 실

1) This work was supported by the Institute of Basic Science Grant 2003 in Inje University.
2) Corresponding Author. Associate Professor, Department of Data Science, Institute of Basic Sciences, Inje University, Kimhae 621-749, Korea.
E-mail : skh@stat.inje.ac.kr
3) Professor, Department of Data Science, Institute of Basic Sciences, Inje University, Kimhae 621-749, Korea.
E-mail : cho@stat.inje.ac.kr

용화되는데 많은 어려움이 있는 실정이다.

이러한 문제를 해결하기 위해서 LS-SVM이 개발되어 최근에 많은 연구가 진행 중이다 (Suykens 등(1999, 2000, 2001), Van Gestel 등(2001)). 많은 논문에서 밝혔듯이 LS-SVM의 수행능력이 SVM에 비해 뒤지지 않을 뿐 아니라 QP 문제를 LP(Linear Programming)문제로 해결하여 훈련시간을 획기적으로 줄이는 방법으로 각광을 받고 있다.

LS-SVM 이나 SVM이 많은 연구에서 좋은 평가를 받는 가장 큰 이유는 아주 큰 데이터를 대표할 수 있는 데이터를 추출하여 차후의 추론에 사용할 수 있다는 것이다. 이러한 분류나 추정에서 중요한 역할을 하는 자료를 SV라 한다. 이렇게 함으로써 계산 할 때 필요한 저장공간을 줄일 수 있고 나아가 계산시간을 줄일 수 있다. SVM에서는 이 SV가 훈련과정에서 자동적으로 정해지지만 LS-SVM에서는 각 데이터의 중요도를 고려하여 사용자가 주관적으로 결정한다. Suykens 등(2000)은 이러한 중요도를 결정하는 과정에서 오직 잔차의 크기만을 고려하였다.

본 연구에서는 LS-SVM에서 SV를 결정할 때 잔차의 크기뿐만 아니라 분산도 고려하여 선택하는 방법을 제안한다. 2절에서는 LS-SVM을 소개하고, 3절에서는 SV를 구하는 기존의 알고리즘과 제안된 알고리즘을 소개하고 모의실험을 통해 제안된 방법의 우수성을 살펴본다.

## 2. LS-SVM

다음과 같은 훈련용자료(training data set) $\{\boldsymbol{x}_k, y_k\}$, $k=1,\ldots,N$ 이 주어 졌다고 하자. 여기에서 $\boldsymbol{x}_k \in R^n$ 이고 $y_k \in R$ 이다. 우리가 추정하고자 하는 비선형함수는 다음과 같이 표현된다고 하자.

$$y_k = f(\boldsymbol{x}_k) + \varepsilon$$

여기에서

$$f(x) = \boldsymbol{w}^T \psi(\boldsymbol{x}) + b \tag{1}$$

이고, $\varepsilon$은 서로 독립이고 평균이 0인 분포를 따르는 확률변수이다. 그리고 $\psi(\cdot): R^n \rightarrow R^m$는 입력공간에서 높은 차원의 특징공간(feature space)으로의 함수이다. (1)의 $f(\boldsymbol{x})$를 추정하기 위하여 다음과 같은 최적화문제를 고려한다.

$$\min_{w,b,e} T(\boldsymbol{w}, \boldsymbol{e}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{\gamma}{2}\sum_{i=1}^{N} e_k^2 \tag{2}$$

$$\textit{subject to} \quad y_k = \boldsymbol{w}^T\phi(\boldsymbol{x}_k) + b + e_k, \quad k=1,\cdots,N.$$

이 식은 정규화 항(regularization term)과 제곱오차항을 가지고 있어 상수항( $b$)가 있는 커널능형회귀(kernel ridge regression)로 해석이 되기도 한다. 위의 최적화 문제를 라그랑제함수(Lagrangian)로 표현하면

$$\Lambda(\boldsymbol{w}, b, \boldsymbol{e}; \boldsymbol{\alpha}) = T(\boldsymbol{w}, \boldsymbol{e}) - \sum_{i=1}^{N} \alpha_k \{\boldsymbol{w}^T\phi(\boldsymbol{x}_k) + b + e_k - y_k\} \tag{3}$$

이 된다. 여기에서 $\alpha_k$는 라그랑제 배수(Lagrange multiplier)이다. (3)으로부터 다음과 같은 조건을 유도 할 수 있다.

$$\begin{cases} \frac{\partial \Lambda}{\partial \boldsymbol{w}} = 0 \rightarrow \boldsymbol{w} = \sum_{k=1}^{N} \alpha_k \psi(\boldsymbol{x}_k) \\ \frac{\partial \Lambda}{\partial b} = 0 \rightarrow \sum_{k=1}^{N} \alpha_k = 0 \\ \frac{\partial \Lambda}{\partial e_k} = 0 \rightarrow \alpha_k = \gamma e_k, \qquad k=1,\ldots,N \\ \frac{\partial \Lambda}{\partial \alpha_k} = 0 \rightarrow \boldsymbol{w}^T \psi(\boldsymbol{x}_k) + b + e_k - y_k = 0, \quad k=1,\ldots,N \end{cases} \tag{4}$$

이 조건을 만족하는 해는 다음과 같은 선형식으로 구할 수 있다.

$$\begin{pmatrix} 0 & 1^T \\ 1 & \Omega + \gamma^{-1} I \end{pmatrix} \begin{pmatrix} b \\ \boldsymbol{\alpha} \end{pmatrix} = \begin{pmatrix} 0 \\ \boldsymbol{y} \end{pmatrix} \tag{5}$$

여기에서 $\boldsymbol{y}^T = (y_1, \ldots, y_N)$, $1^T = (1, \ldots, 1)$, $\boldsymbol{\alpha}^T = (\alpha_1, \ldots, \alpha_N)$, $0^T = (0, \ldots, 0)$이고 $\Omega$는 $N \times N$ 행렬인데 $(k, l)$번째 원소 $\Omega_{kl}$는 Mercer 정리에 의해 다음과 같이 커널함수 $K$로 표현할 수 있다.

$$\begin{aligned} \Omega_{kl} &= \psi(\boldsymbol{x}_k)^T \psi(\boldsymbol{x}_l) \\ &= K(\boldsymbol{x}_k, \boldsymbol{x}_l). \end{aligned}$$

많이 사용되고 있는 커널함수로는 아래와 같은 것들이 있다.

$$K(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T \boldsymbol{y} \text{ (선형커널함수)}$$

$$K(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^T \boldsymbol{y} + 1)^d \text{ (차수가 } d\text{인 다항커널함수)}$$

$$K(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|_2^2}{\sigma^2}\right) \text{(RBF 커널함수)}$$

(4)와 (5)로부터 $f(\boldsymbol{x})$의 추정치 $\hat{f}$를 다음과 같이 얻을 수 있다.

$$\hat{f}(\boldsymbol{x}) = \sum_{k=1}^{N} \alpha_k K(\boldsymbol{x}, \boldsymbol{x}_k) + b \tag{6}$$

여기에서 $\alpha_k$와 $b$는 (5)의 해이다. $\alpha_k$는 $\hat{f}(\boldsymbol{x})$에서 $\boldsymbol{x}_k$의 중요도를 나타내는 가중치로 해석이 가능하다. 그런데 이는 (4)에서 알 수 있듯이 오차의 크기 $e_k$에 비례한다. 즉, 오차의 크기가 큰 값에 해당하는 입력값이 바로 SV인데 이는 함수의 추정에 중요한 역할을 하는 것으로 알려져 있다. 이러한 사실을 이용하여 Suykens 등(2000)은 SV를 구하는 알고리즘을 제안하였다. 이 알고리즘을 다음 절에서 소개한다.

## 3. SV를 구하는 알고리즘과 모의실험

$\alpha_k$는 $\hat{f}(x)$에서 $x_k$의 중요도를 나타내는 측도로 이용할 수 있다. 그래서 Suykens(2000)은 $|\alpha_k|$의 크기가 큰 것에 대응하는 $x_k$를 SV로 선택하는 방법을 아래와 같이 제안하였다.

1. 전체 자료를 이용하여 (6)식의 $\alpha_1, \ldots, \alpha_N$ 을 구한다.

2. $|\alpha_1|,\ldots,|\alpha_M|$ 을 크기순으로 나열하여 작은 값을 가지는 $\beta$%(주어진 비율, 5% ~ 20%)에 해당하는 자료를 제외한 새로운 자료를 얻는다.
3. 새로운 자료를 전체자료로 놓고 처음의 전체 자료 중 약 $k$%(1% ~10%)가 남을 때까지 단계 1, 2를 반복한다. 이렇게 남겨진 자료가 SV이다.

이 알고리즘은 $|\alpha_k|$의 크기가 큰 것에 대응하는 $x_k$를 SV로 선택하는 방법이다. 그러나 이 알고리즘은 등분산모형에서는 결과가 상당히 좋은 것으로 판명이 되었다. 그러나 이분산 모형(heteroscedastic model)에서 이 방법을 사용하게 된다면 분산이 작은 부분에서는 잔차가 적어질 확률이 크다. 그래서 이 부분에서의 자료가 SV로 선택되어질 확률도 상당히 적어질 것이다. 이렇게 된다면 SV만을 가지고 추정을 하면 그 결과가 편향될 것으로 생각된다. 그림 1.에서 이와 같은 결과를 보이고 있다. 크기가 500인 자료는

$$y=\sin(x)+e \tag{7}$$

를 통하여 얻었다. 여기에서 $x$는 $[-\pi,\pi]$에서 등간격 값을 가지도록 하였고, $e=0.3(\sin(x-\pi/2))z$ 인데, $z$는 표준정규분포를 따르는 확률변수다. 그리고 본 연구에서는 RBF 커널함수

$$K(\boldsymbol{x},\boldsymbol{y})=\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|_2^2}{\sigma^2}\right)$$

를 사용하였다. 그리고 실험에 필요한 커널모수 $\sigma$와 정규화 모수 $\gamma$는 10-fold 교차타당성방법(cross validation)을 이용하여 구하였다. 그림 1.에서 자료는 점으로 표시되어있고, 전체자료의 20%인 SV는 작은 원으로 표시되어있다. 그리고 $y=\sin(x)$는 실선으로, SV만으로 추정한 값은 점선으로 나타나 있다. 위에서 언급하였듯이 분산이 작은 부분, $x=\pm\frac{\pi}{2}$ 근방에서는 SV가 거의 없고, 이 부분에서 멀어질수록 SV가 많아짐을 알 수 있다. 그 결과 $x=-\frac{\pi}{2}$ 부분에서의 추정값이 편향되어 있음을 알 수 있다. 그러나 $x=\frac{\pi}{2}$ 부분에서는 추정이 잘 되는 것으로 나타났는데, 이는 이 주위의 SV가 우연히도 적절하게 잘 잡혀졌기 때문이다.

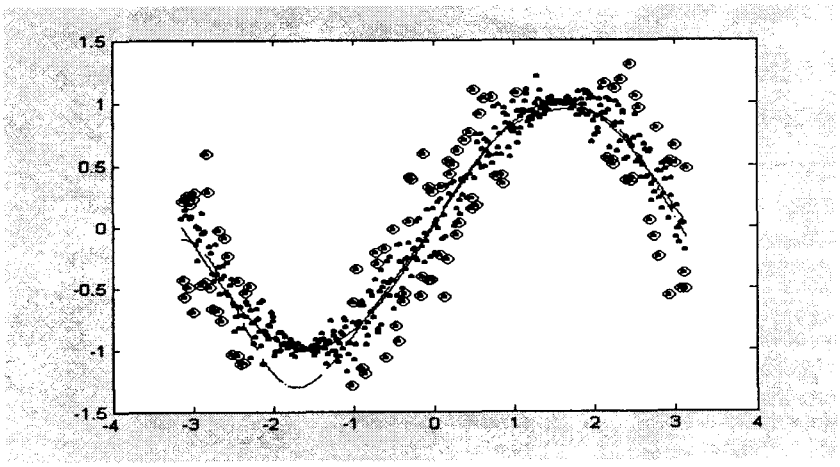


그림 1. 이분산 모형에서 기존의 방법으로 구한 SV와 이를 이용한 추정

이와 같이 이분산모형에서 SV를 구할 때는 잔차의 크기와 분산을 함께 고려해 주어야 할 것이다. 그래서 본 논문에서는 Suykens 등(2000)의 방법에서 $\alpha_i$ 대신에 $\sigma_i\alpha_i$를 사용 할 것을 제안한다. 여기에서 $\sigma_i$는 오차 $\epsilon$의 분산의 국소 추정치로써 어떠한 것이라도 사용이 가능한데 본 연구의 목적이 이것의 추정에 있는 것이 아니므로 간단한 형태의 추정치

$$\sigma_i = \frac{1}{2u-p-1} \sum_{k\in(i-u,\, i+u)} (\hat{f}(x_k)-y_k)^2$$

를 사용한다. 여기에서 $p$는 $\hat{f}(x)$의 자유도인데 Vapnik(1998)을 참고하여

$$p = \sum_{k=1}^{n} \frac{\lambda_k}{\lambda_k+\gamma}$$

를 이용하여 구하였고 $u=20$을 사용하였다. 여기에서 $\lambda_k$는 $K^TK$의 고유값이고 $\gamma$는 정규화 모수이다.

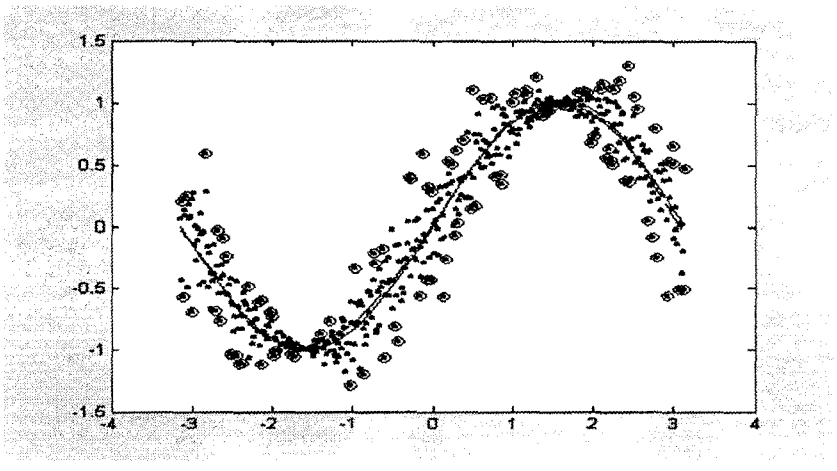


그림 2. 이분산 모형에서 제안된 방법으로 구한 SV와 이를 이용한 추정

제안된 방법으로 구한 SV와 구해진 SV를 이용한 추정을 그림 2.에 나타내었다. 그림 1.과 비교해 보면 더 좋은 결과를 나타냄을 알 수 있다. 특히 $x=\pm\frac{\pi}{2}$ 부근에서도 SV가 선택이 되었고

| | 등분산성모형 | | 이분산성모형 | |
|---|---|---|---|---|
| | AMSE | SDMSE | AMSE | SDMSE |
| 모든자료 | 0.0017 | 0.00075 | 0.00075 | 0.00046 |
| 제안된 방법 | 0.0039 | 0.00201 | 0.0016 | 0.0011 |
| 기존의 방법 | 0.0048 | 0.00314 | 0.01091 | 0.0099 |

표 1. 여러 방법에 의한 추정치의 모의실험 결과

이로 인해 이 근방에서 편의도 없음을 알 수 있다. 반복된 실험에서도 좋은 결과를 보이는지를 알

아보기 위하여 100번의 반복 된 실험을 통하여 MSE의 평균(AMSE)과 표준편차(SDMSE)를 구하여 보았다. 표 1.에서 실험결과를 볼 수 있는데, 모든 자료를 이용한 추정, 기존의 방법으로 구한 SV를 이용한 추정 그리고 제안된 방법으로 구한 SV를 이용한 추정의 결과를 나타내었다. 그리고 (7)식의 모형에서 오차항이 $e=0.3z$ 인 등분산성 모형에서의 수행능력도 표 1.에 나타내었다. 모든 자료를 사용한 추정치는 모든 모형에서 좋은 결과를 보인다는 것이 이 표에서 확인되었다. 그리고 제안된 방법은 등분산성모형과 이분산성의 모형 모두에서 기존의 방법보다 더 좋은 것으로 나타났다. 특히 예상한 것과 같이 이분산성모형에서 더욱더 좋은 결과를 나타내었다. 그러나 모든 자료를 이용한 추정치에 비해서는 좋지 않은 것으로 나타나는데 20%의 자료를 이용한 결과라는 것을 감안하면 실망스러운 것은 아니라고 생각한다.

## 4. 참고문헌

[1] Bishop C. M.(1995), *Neural networks for pattern recognition*, Oxford University Press.

[2] Cherkassky V., Mulier F.(1998), *Learning from data: concept, theory and method*, John Wiley and Sons.

[3] Cristianini N, Shawe-Taylor J.(2000), *An Introduction to Support Vector Machines*, Cambridge University Press.

[4] Haykin S.(1994), *Neural Networks: a Comprehensive Foundation*, Macmillan College Publishing Company: Englewood Cliffs.

[5] Suykens J.A.K., Vandewalle J.(1999), Least squares support vector machine classifiers, *Neural Processing Letters*, Vol.9, No.3, pp293-300.

[6] Suykens J.A.K., Lukas L., Vandewalle J.(2000), Sparce approximation using least squares support vector machines, *IEEE International Symposium on Circuits and Systems* (ISCAS 2000), pp.II757-II760, Geneva, Switzerland.

[7] Suykens J.A.K., Vandewalle J, De Moor B.(2001), Optimal control by least squares support vector machines, *Neural Networks*, Vol.14, No.1, pp23-35.

[8] Van Gestel T., Suykens J.A.K., Baesens B., Viaene S., Vanthienen J., Dedene G., De Moor B.(2001b), Benchmarking least squares support vector machine classifiers, *Internal Report* 00-37, *ESAT-SISTA, K.U. Leuven.*

[9] Vapnik V.(1998), *Statistical learning theory*, John Wiley, New-York.

[ 2003년 8월 접수, 2003년 10월 채택 ]