Journal of Korean
Data & Information Science Society
2003, Vol. 14, No.3 pp. 451~459

Confidence Interval Estimation Using SV in LS-SVM[1)]

Kyung Ha Seok[2)]

Abstract

The present paper suggests a method to estimate confidence interval using SV(Support Vector) in LS-SVM(Least-Squares Support Vector Machine). To get the proposed method we used the fact that the values of the hessian matrix obtained by full data set and SV are not different significantly. Since the suggested method implement only SV, a part of full data, we can save computing time and memory space. Through simulation study we justified the proposed method.

Keywords : confidence interval, SV, LS-SVM

1. 서론

최근 수십년 동안 신경망은 많은 분야에서 좋은 분석 그리고 추정 방법으로 각광을 받아 왔다(Bishop(1995), Cherkassky 등(1998), Haykin(1994)). 그럼에도 불구하고 은닉층과 은닉 노드의 수를 결정 할 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 SV로 해결이 된다.

많은 논문에서 여러 가지의 자료를 통해 SVM의 우월성을 증명하고 있다. 그러나 SVM을 훈련시키기 위해서는 QP(Quadratic Programming)문제를 해결해야 하는 단점이 있다. 이는 계산에 많은 시간과 공간을 요구한다. 실제로 규모가 큰 자료에서는 SVM을 훈련하는데 2~3일 정도가 소요되는 것으로 알려져 있다. 이러한 단점 때문에

---

1) This work was supported by the 2002 Inje University research grant.
2) 경남 김해시 어방동 인제대학교 데이터정보학과, 인제대학교 기초과학연구소

SVM이 실용화되는데 많은 어려움이 있는 실정이다.

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

(LS-)SVM의 가장 큰 장점 중 하나는 추정이나 분류를 할 때 모든 데이터를 사용하지 않고 필요한 자료만 이용을 한다는 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.

비선형 회귀함수를 추정하여 여러 가지 추론을 하는 과정에서 신뢰구간은 중요한 역할을 한다. 위에서 언급한 바와 같은 이유로 SV만으로 신뢰구간을 구하는 방법은 여러모로 필요한 방법이 되겠다. 그러나 아직까지는 여기에 대한 연구가 되지 않은 실정이다. 이에 본 연구는 SV만을 이용하여 신뢰구간을 구하는 방법을 제안한다.

제 2절에서는 LS-SVM에 대한 간략한 소개를 하고, 3절에서는 제안된 방법의 설명을 하고 4절에서는 모의실험을 통한 제안된 방법의 타당성을 입증한다.

2. LS-SVM

다음과 같은 훈련용자료(training data set) $\{\boldsymbol{x}_k, y_k\}$, $k=1,\cdots,N$ 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

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

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

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

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

이 식은 정규화 항(regularization term)과 제곱오차항을 가지고 있어 상수항(b)가 있는 커널능형회귀(kernel ridge regression)로 해석이 되기도 한다(Jianhua Xu 등(2001)). 위의 최적화 문제를 라그랑제함수(Lagrangian)로 표현하면

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

이 된다. 여기에서 α_k는 라그랑제 배수(Lagrange multiplier)이다. (3)으로부터 다음과

같은 조건을 유도 할 수 있다.

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

이 조건을 만족하는 해는 다음과 같은 선형식으로 구할 수 있다.

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

여기에서 $\boldsymbol{y}^T = (y_1, \cdots, y_N)$, $\mathbf{1}^T = (1, \cdots, 1)$, $\boldsymbol{\alpha}^T = (\alpha_1, \cdots, \alpha_N)$, $\mathbf{0}^T = (0, \cdots, 0)$이고 Ω는 $N \times N$ 행렬인데 (k, l)번째 원소 Ω_{kl}는 Mercer 정리에 의해 다음과 같이 커널함수 K로 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.

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

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

$K(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x}^T - \boldsymbol{y}\|_2^2}{\sigma^2}\right)$ (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 α_k와 b는 (5)의 해이다. α_k는 $f(\boldsymbol{x})$에서 $\boldsymbol{x}_k$의 중요도를 나타내는 가중치로 해석이 가능하다. 그런데 이는 (4)에서 알 수 있듯이 오차의 크기 e_k에 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.

3. SV를 이용한 신뢰구간

비선형 회귀함수의 신뢰구간은 Chryssolouris 등(1996)에서 연구되었다. 이 결과를 응용하면 입력값 $\boldsymbol{x}_0$에 대한 추정값 $\hat{f}(\boldsymbol{x}_0)$를 이용한 $100(1-\alpha)\%$ 신뢰구간은

$\hat{f}(\boldsymbol{x}_0)\pm c$로 주어진다. c는

$$c = t_{N-p}^{\alpha/2} s (1 + \boldsymbol{J}_0^T (\boldsymbol{F}^T \boldsymbol{F} + \gamma^{-1} I)^{-1} \boldsymbol{F}^T \boldsymbol{F} (\boldsymbol{F}^T \boldsymbol{F} + \gamma^{-1} I)^{-1} \boldsymbol{J}_0)^{1/2} \quad (7)$$

이다. 여기에서

$$s^2 = \sum_{k=1}^{N} (y_i - \hat{f}(\boldsymbol{x}_k))^2 / (N-p)$$

$$\boldsymbol{J}_0 = (K(\boldsymbol{x}, \boldsymbol{x}_1), K(\boldsymbol{x}, \boldsymbol{x}_2), \cdots, K(\boldsymbol{x}, \boldsymbol{x}_N))^T$$

$$\boldsymbol{F} = \begin{bmatrix} K(\boldsymbol{x}_1, \boldsymbol{x}_1) & K(\boldsymbol{x}_1, \boldsymbol{x}_2) & \cdots & K(\boldsymbol{x}_1, \boldsymbol{x}_N) \\ K(\boldsymbol{x}_2, \boldsymbol{x}_1) & K(\boldsymbol{x}_2, \boldsymbol{x}_2) & \cdots & K(\boldsymbol{x}_2, \boldsymbol{x}_N) \\ \vdots & \vdots & \vdots & \vdots \\ K(\boldsymbol{x}_N, \boldsymbol{x}_2) & K(\boldsymbol{x}_N, \boldsymbol{x}_2) & \cdots & K(\boldsymbol{x}_N, \boldsymbol{x}_N) \end{bmatrix}$$

이고, p는 모수의 수이다. 이 신뢰구간은 LS-SVM에 적용이 잘 된다(Hwang 등 (2002)). 만약 자료의 크기가 크다면 (7)식을 계산하기 위한 시간이나 공간이 많이 필요하다. 그래서 a개의 SV, $(\boldsymbol{x}_1^{SV}, \boldsymbol{x}_2^{SV}, \cdots, \boldsymbol{x}_a^{SV})$를 구할 수 있다면 이를 이용하여 신뢰구간을 구하면 계산이 훨씬 용이 할 것이다.

SV를 이용한 추정값 $\hat{f}_{SV}(\boldsymbol{x}_0)$가 $\hat{f}(\boldsymbol{x}_0)$에 가까운 값을 가진다는 것이 Syukens 등 (2001)에서 보여 주었다. 그러나 신뢰구간을 구하는 작업도 매번 많은 계산을 거쳐야 하기 때문에 SV를 이용할 수 있으면 적은 저장공간으로 빠른 계산을 할 수 있을 것이다. 그러므로 SV를 이용하여 신뢰구간을 구할 수 있는 방법을 제시한다.

(7)식의 c는 $t-$값, s 그리고

$$h(\boldsymbol{x}_0) = (1 + \boldsymbol{J}_0^T (\boldsymbol{F}^T \boldsymbol{F} + \gamma^{-1} I)^{-1} \boldsymbol{F}^T \boldsymbol{F} (\boldsymbol{F}^T \boldsymbol{F} + \gamma^{-1} I)^{-1} \boldsymbol{J}_0)^{1/2}$$

의 3부분으로 구성되었다. 먼저 s를 고려하면 SV는 오차가 큰 값에 해당하는 값이므로 이를 이용한 추정치의 s는 아주 커진다. 그 뿐만 아니라 전체자료를 이용한 s를 이용할 수 있으므로 여기에 대해서는 SV로 구한 값을 고려 할 필요가 없다. 그러므로 $t-$값도 그대로 사용하면 된다.

한편, h는 신뢰구간을 구할 때마다 계산해야 하는데 이때 계산시간을 절약하기 위하여 SV로부터 구해진 h_{SV}를 사용한 근사치를 구하는 방법을 아래와 같이 제안한다.

① 먼저 경계에 위치하지 않은 임의의 $\boldsymbol{x}$에 대하여 h와 h_{SV}를 계산한다.

② $b = h/h_{SV}$를 계산한다.

③ 신뢰구간을 $\hat{f}_{SV}(\boldsymbol{x}_0)\pm c_{SV}$, $c_{sv} = c \times b \times h_{SV}$ 로 계산한다.

이러한 방법을 제안 할 수 있는 이유는 오차가 등분산성을 만족한다는 가정과 전체자료와 SV를 가지고 만든 헤시안행렬(Hessian matrix)의 값이 크게 다르지 않기 때문이다.

다음 절에서는 제안된 방법의 정당성을 모의 실험을 통하여 입증한다.

4. 모의실험

제안된 방법의 정당성을 입증하기 위하여 모의실험을 하였다. 이 실험을 위하여 sine 함수를 이용하였다.

$$y=\sin(x)+\varepsilon \quad , \ -\pi<x<\pi.$$

여기에서 $\varepsilon\sim N(0,0.3^2)$의 분포를 따른다. 본 연구를 위하여 RBF 커널함수

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

를 사용하였다. 그리고 실험에 필요한 커널모수 σ와 정규화 모수 γ는 10-fold 교차타당성방법(cross validation)을 이용하여 구하였다. 그리고 (7)식의 p는 Vapnik(1998)을 참고하여

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

를 이용하여 구하였다. 여기에서 λ_k는 K^TK의 고유값이고 γ는 정규화 모수이다.

먼저 크기가 500인 표본을 뽑아 $\hat{f}(x)$를 구하고, 48개의 SV를 이용한 $\widehat{f_{SV}}(x)$를 구하여 그림 1에 나타내었다. 그림 1에서 보듯이 전체의 약 10%의 자료, SV를 이용한 추정(dotted line)이 크기가 $n=500$인 전체자료를 이용한 추정(dashed line)과 비슷한 결과를 보임을 알 수 있다. 이 그림과 그림 2, 3에서 점은 자료를 원은 SV를 나타낸다.

다음으로 신뢰구간을 구하여 보았다. 여기에서는 (7)식을 이용한 신뢰구간을 구하여 그림 2에 나타내었다.

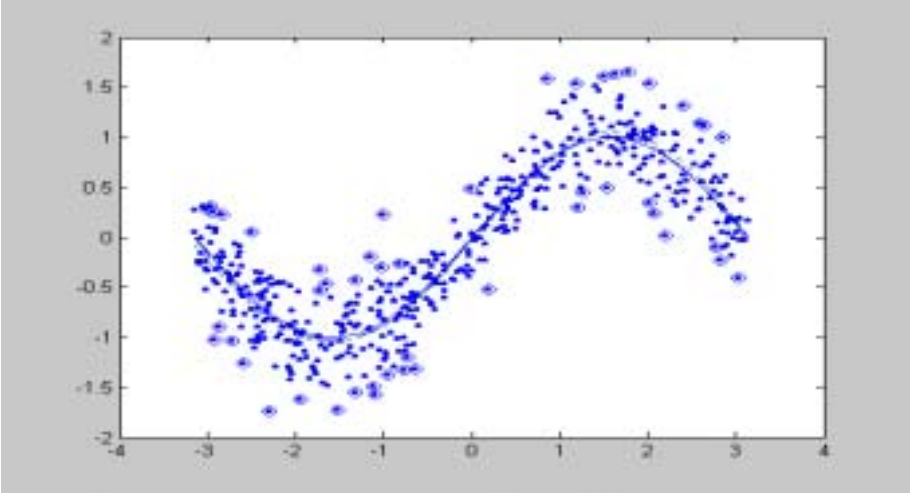

그림 1. SV를 이용한 sine함수추정.

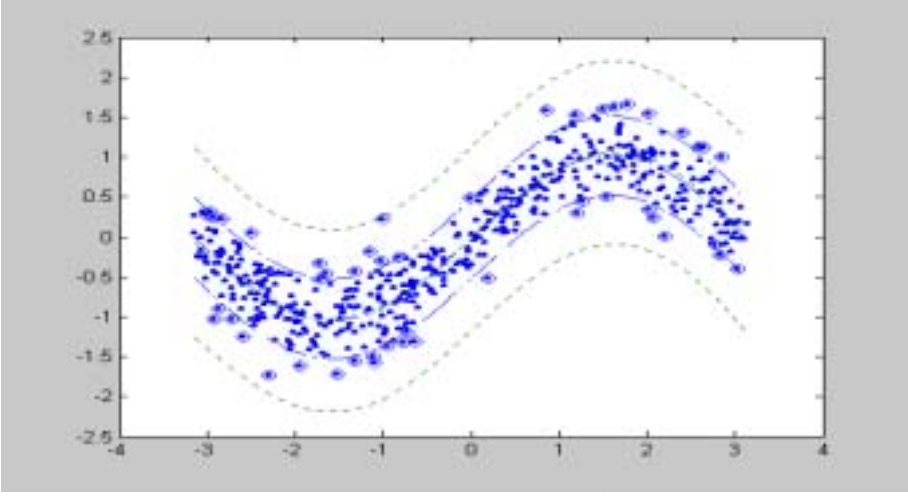

그림 2. 모든자료를 이용한 신뢰구간과 SV를 이용한 신뢰구간

그림 2에서 SV를 이용한 신뢰구간(dotted line)의 폭이 모든 자료를 이용한 신뢰구간(dashed line)보다 상당히 커짐을 알 수 있다. 이는 앞에서 언급한 바와 같이 s의 값이 많이 커지기 때문이다.

제안된 방법으로 구해진 신뢰구간(dotted line)과 모든 자료를 이용한 신뢰구간(dashed line)을 구하여 그림 3에 나타내었다. 이 그림에서 두 신뢰구간이 상당히 유사함을 알 수 있다. 그리고 그림에서 나타나는 신뢰구간의 차이는 $\widehat{f}(x)$와 $\widehat{f_{SV}}(x)$의 차이에서 기인한다는 것을 알 수 있었다. 그러므로 제안된 신뢰구간의 수행능력(performance)이 좋은 것으로 예상을 할 수 있다. 반복된 실험에서도 이와 같은 결과가 나오는지를 알아보기 위하여 이러한 실험을 100번 반복하여 제안 된 방법의 정당성을 살펴보았다. 신뢰구간의 중심은 차이가 거의 없으므로 신뢰구간의 폭의 상대적 차이(relative difference)를 통하여 두 신뢰구간을 비교하여 보았다. 이 결과를 히스토그램으로 그림 4에 나타내었다. 상대적 차이의 절대값이

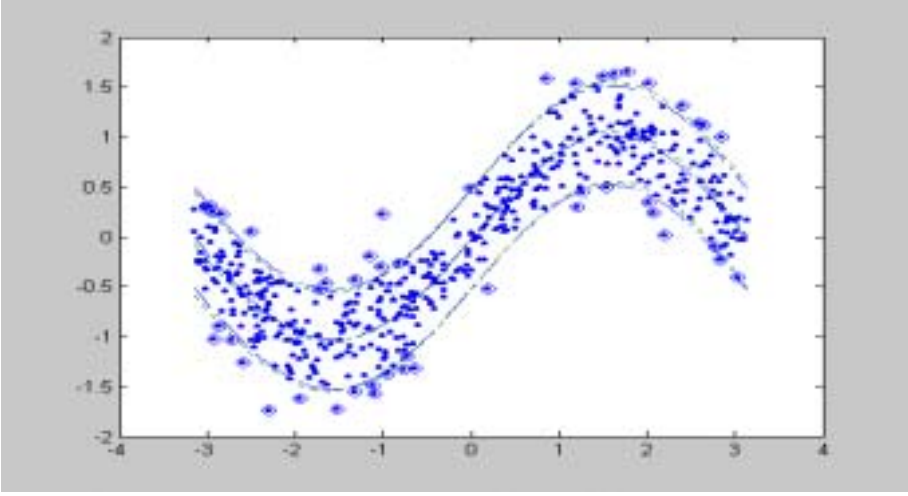

그림 3. 모든 자료를 이용한 신뢰구간(dashed line)과 SV를 이용한 제안된 신뢰구간(dotted line)

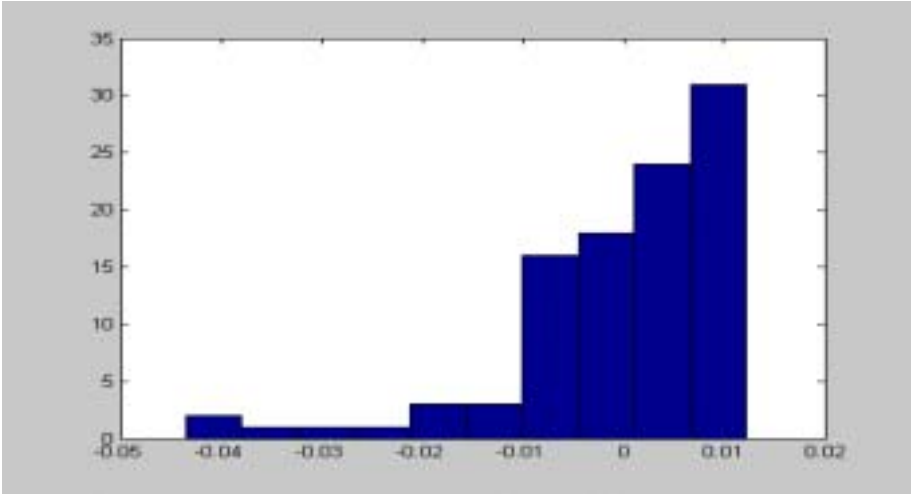

그림 4. 모든 자료를 이용한 신뢰구간과 제안된 방법으로 구한 신뢰구간의 크기의 상대적 차이의 히스토그램

0.02이하인 경우가 95% 정도가 되었고 모든 경우에서 (-0.04, 0.015)사이에 위치 해 있었다. 이를 통해 제안된 방법으로 구한 신뢰구간은 모든 자료를 이용한 신뢰구간의 근사치로 사용 될 수 있음을 입증하였다.

참고문헌

1. Bishop C. M.(1995), *Neural networks for pattern recognition*, Oxford University Press.
2. C. Hwnag, K. Seok, D. Cho(2002), A prediction interval estimation method for KMSE. submitted to *Computers and Chemical Engineering*.
3. Cherkassky V., Mulier F.(1998), *Learning from data*: *concept, theory and method*, John Wiley and Sons.
4. Chryssolouris, G., Lee, M., Ramsey, A.,(1996), Confidence interval prediction for neural networks, *IEEE Trans. Neural Networks* 7, 1, 229-232.
5. Cristianini N, Shawe-Taylor J.(2000), *An Introduction to Support Vector Machines*, Cambridge University Press.
6. Haykin S.(1994), *Neural Networks*: *a Comprehensive Foundation*, Macmillan College Publishing Company: Englewood Cliffs.
7. Jianhua Xu, Xuegong Zhang, and Yanda Li(2001), Kernel MSE algorithm : A unified framework for KFD, LS-SVM, Proceedings of IJCNN'01, 2: 1486-1491.
8. Suyken J.A.K., Vandewalle J.(1999a), Least squares support vector machine classifiers, *Neural Processing Letters*, Vol.9, No.3, pp293-300.
9. Suyken J.A.K., Lukas L., Van Dooren P., De Moor B., Vandewalle J.(1999b), Least squares support vector machine classifiers: a large scale algorithm, *European Conference on Circuit Theory and Design*, (ECCTD'99), pp839-842, Stresa Italy.
10. Suyken J.A.K., Lukas L., Vandewalle J.(2000), Sparse approximation using least squares support vector machines, *IEEE International Symposium on Circuits and Systems*(ISCAS 2000), pp.II757-II760, Geneva, Switzerland.
11. Suyken 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.
12. Van Gestel T., Suyken J.A.K., De Moor B.,Vandewalle J.(2001a), Automatic relevance determination for least squares support vector machine regression, *9th European Symposium on Artificial Neural Networks*(ESANN 2001), pp.13-18, Burges Belgium.

13. Van Gestel T., Suyken 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*.
14. Van Gestel T., Suyken J.A.K., Baestaens D., Lambrechts A., Lanckriet G., Vandaele B., De Moor B., Vandewalle J.(2001c), Financial time series prediction using least squares support vector machines within the evidence framework, *IEEE Transactions on Neural Networks*.
15. Vapnik V.(1998), *Statistical learning theory*, John Wiley, New-York

[2003년 3월 접수, 2003년 6월 채택]