Journal of Korean
Data & Information Science Society
2003, Vol. 14, No.1 pp 67~74

Support Vector Median Regression[1)]

Changha Hwang[2)]

Abstract

Median regression analysis has robustness properties which make it an attractive alternative to regression based on the mean. Support vector machine (SVM) is used widely in real-world regression tasks. In this paper, we propose a new SV median regression based on check function. And we illustrate how this proposed SVM performs and compare this with the SVM based on absolute deviation loss function.

Keywords : Median regression, neural network, support vector machine

1. 서론

SVM은 원래 분류(classification)를 위해 Vapnik과 공동연구자들에 의해 개발되어 문자인식, 얼굴인식, 생물정보학 등의 응용분야에 적용되어 좋은 결과들을 보여주고 있다. 최근 분류를 위한 SVM 이론이 회귀분석 및 함수추정으로 확장되어 많이 활용되고 있다. SVM은 투사지향(projection pursuit) 알고리즘 및 신경망과 함께 주로 두 개 이상의 독립변수를 갖는 비선형 함수의 추정을 위해 사용되는 방법이다. SVM은 VC-이론을 이용하여 예측(prediction)을 잘하고 벌칙항(penalty term)을 사용하여 과대적합(overfitting)을 피하려는데 초점을 두고 있다. 여기서 VC는 이론 창시자들의 이름인 Vapnik과 Chervonenkis 첫 글자를 따서 표현된 것이다. SVM은 볼록함수(convex function)를 최소화하여 학습이 진행되기 때문에 신경망과는 달리 유일한 해를 구할 수 있는 장점이 있다. SVM 회귀분석은 ε-무감각(insensitive) 오차함수를 사용하고 벌칙항을 이용하여 과대적합을 피하기 때문에 이상치에 민감하지 않고 그 영향을 최소화시키는 로버스트 회귀분석 방법이다. 자세한 내용을 위해 Gunn(1998), Smola & Schoelkopf(1998) 및 Vapnik(1995,1998) 등을 참조하라. 한편 로버스트 회귀분석을 위해 주로 사용되는 중위수 회귀분석(median

---

1) This research was supported by the Catholic University of Daegu research grants in 2003.

2) Associate Professor, Dept. of Statistical Information, Catholic University of Daegu.
E-mail : chhwang@cataegu.ac.kr

regression)은 기본적으로 절대편향(absolute deviation) 오차함수를 사용하며, 이 오차함수는 $\varepsilon=0$ 인 경우의 ε-무감각 오차함수이므로 SVM의 기본원리를 이용하여 새로운 방법의 중위수 회귀분석을 유도할 수 있다.

본 논문에서는 Koenker & Bassett(1978)) 또는 Koenker & Hallock(2001)의 대조함수(check function)를 사용하는 SV 중위수 회귀분석을 제안하고 절대편향 오차함수를 사용하는 SV 중위수 회귀분석과 비교 · 분석한다. 그리고 계량경제학에서 현재 로버스트 회귀분석을 위해 많이 사용되고 있는 신경망 중위수 회귀분석과도 비교한다.

2. 절대편향 오차함수 기반 SV 중위수 회귀분석

통계학에서는 중위수 회귀분석을 위해 전통적으로 절대편향 오차함수를 사용한다. 그런데 절대편향 오차함수는 SVM 회귀분석을 위해 주로 사용되는 ε-무감각 오차함수에서 $\varepsilon=0$ 인 경우이다. 본 절에서는 절대편향 오차함수를 사용하는 SVM 회귀분석에 대해 간단히 생각해 보며, 이 회귀분석을 절대편향 오차함수 기반 SV 중위수 회귀분석이라 부르고자 한다. 먼저 선형 중위수 회귀분석을 위한 SVM을 간단하게 알아본다. 자세한 내용은 참고문헌 Gunn(1998), Smola & Schoelkopf(1998) 및 Vapnik(1995, 1998)에 설명되어 있다.

훈련자료 $\{(x_i, y_i),\ i=1,\cdots,n\}\subset \mathfrak{X}\times R$ 가 주어졌다고 가정한다. 여기서 $\mathfrak{X}$는 입력벡터공간 R^d를 나타낸다. 우리들의 목표는 모든 훈련자료에 대해서 가능한 작은 크기의 w값을 갖는 함수 $f(x)$를 찾는 것이다. 다음과 같은 함수형태를 생각한다.

$$f(x) = w^t x + b \text{ with } w\in\mathfrak{X},\ b\in R$$

여기서 윗첨자 t는 행렬의 전치를 나타낸다. 이 이론은 가중치 감소(weight decay)를 사용하여 회귀함수를 추정하는 다층전방향 신경망에도 활용되고 있다. 이를 위한 한 가지 방법은 유클리드 놈 $\|w\|^2$을 최소화하는 것이다. 우리는 이 문제를 다음과 같은 볼록 최적화(convex optimization) 문제로 간주할 수 있다.

$$\text{minimize} \quad \frac{1}{2}\|w\|^2,$$
$$\text{subject to} \quad y_i - w^t x_i - b \le 0 \quad \text{and} \quad w^t x_i + b - y_i \le 0$$

여기서 기본가정은 볼록 최적화 문제가 해결 가능하다는 것이다. 그러나 가끔 이 가정이 성립되지 않는다. 따라서 이를 해결하기 위해서 새로운 변수 $\xi_i,\ \xi_i^*$를 도입한다. 그러면 Vapnik(1995, 1998)이 제안한 다음과 같은 최적화 문제로 귀착된다.

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

$$\text{subject to} \quad \begin{cases} y_i - w^t x_i - b \le \xi_i \\ w^t x_i + b - y_i \le \xi_i^* \\ \xi_i,\ \xi_i^* \ge 0 \end{cases}$$

벌칙상수 $C>0$는 함수 f의 평평함과 오차합의 균형을 고려하여 결정된다. 여기서 ξ_i, ξ_i^*는 출력에 대한 상계제약(upper constraint)과 하계제약(lower constraint)을 나타낸다.

위의 최적화 문제의 핵심 이론은 다음과 같은 라그랑즈(Lagrange) 함수를 만드는 것이다.

$$L = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}\alpha_i(\xi_i - y_i + w^t x_i + b)$$

$$-\sum_{i=1}^{n}\alpha_i^*(\xi_i^* + y_i - w^t x_i - b) - \sum_{i=1}^{n}(\eta_i\xi_i + \eta_i^*\xi_i^*)$$

여기서 제약조건 α_i, α_i^*, η_i, $\eta_i^* \geq 0$이 성립해야 한다. 따라서 최적화 문제는 다음과 같이 된다.

$$\text{maximize} \quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)x_i^t x_j + \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*),$$

$$\text{subject to} \quad \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \quad \text{and} \quad \alpha_i,\ \alpha_i^* \in [0,\ C]$$

이런 제약조건을 가진 위의 방정식을 푸는 것은 라그랑즈 배수(Lagrange multiplier) α_i, α_i^*를 구하는 것을 의미한다. 추정된 최적의 회귀계수는 다음과 같다.

$$w = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)x_i, \quad b = -\frac{1}{2}w^t[x_r + x_s]$$

즉, 추정된 최적의 선형회귀함수는 다음과 같이 표현된다.

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)x_i^t x + b$$

여기서 x_r과 x_s는 스포트 벡터이다. 스포트 벡터는 라그랑즈 배수 중 0 보다 큰 α_i, α_i^*에 대응되는 입력벡터이다.

이제 비선형 회귀함수 추정을 위한 절대편향 오차함수 기반 SV 중위수 회귀분석에 대해 알아보자. 위의 방법을 비선형 회귀함수 추정의 경우로 확장하고자 한다. 먼저 SVM의 학습에서 자료는 내적(dot product) $x_i^t x_j$를 통해서 함수에 나타난다. 이제 함수 $\Phi: R^d \rightarrow E$를 사용하여 자료를 유클리드 공간 E로 대응시킨다고 가정한다. 그러면 학습 알고리즘은 E에서 내적 $\Phi(x_i)^t\Phi(x_j)$을 통해 자료에 의존한다. 만일 $K(x_i, x_j) = \Phi(x_i)^t\Phi(x_j)$를 만족하는 커널 함수 K가 존재한다면 우리들은 학습 알고리즘에 단지 K만을 사용하고 Φ가 구체적으로 무엇인지를 알 필요가 없다. 일반적으로 많이 사용되는 커널함수는 다음과 같은 다항 커널과 가우시안 커널이다.

$$K(x, y) = (x^t y + 1)^p, \quad K(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}.$$

여기서 p와 σ^2은 커널 모수이다. 커널 방법은 차원문제(curse of dimensionality)를

해결하기 위해서 사용된다. 절대편향 오차함수 기반 비선형 SV 중위수 회귀분석의 해는 다음과 같은 최적화문제를 통하여 구해진다.

$$\text{maximize} \quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(x_i, x_j)+\sum_{i=1}^{n}y_i(\alpha_i-\alpha_i^*)$$

$$\text{subject to} \quad \sum_{i=1}^{n}(\alpha_i-\alpha_i^*)=0 \quad \text{and} \quad \alpha_i, \alpha_i^* \in [0, C]$$

이런 제약조건하에서 위의 방정식을 푸는 것은 라그랑즈 배수 α_i, α_i^*를 구하는 것이며 추정된 최적의 회귀계수는 다음과 같다.

$$w = \sum_{i=1}^{n}(\alpha_i-\alpha_i^*)\Phi(x_i), \quad b = -\frac{1}{2}\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)[K(x_r, x_i), K(x_s, x_i)]$$

즉, 추정된 최적의 비선형 회귀함수는 다음과 같이 표현된다.

$$f(x) = \sum_{i=1}^{n}(\alpha_i-\alpha_i^*)K(x_i, x)+b$$

선형인 경우와의 차이점은 w가 더 이상 정확한 형태로 주어지지 않는다는 것이다. 그러나 w는 내적을 통해서 유일하게 결정된다. 한편 비선형의 경우에 최적화 문제는 입력공간이 아닌 특징공간(feature space)에서 가장 평평한 함수를 찾는 것에 대응된다. 회귀함수 추정 문제에서 두가지 모수, 즉 벌칙상수 C 및 커널모수가 결정되어야 한다.

3. 대조함수 기반 SV 중위수 회귀분석

Koenker & Bassett(1978)에 의해 제안된 선형 중위수 회귀모형은 y_i가 입력변수 벡터 x_i와 선형관계를 가진다는 가정하에서 목적함수

$$\sum_{i=1}^{n}\rho(y_i - w^t x_i - b)$$

를 최소화 하는 $f(x) = w^t x + b$를 구하는 것이다. 여기서 $\rho(\cdot)$는 대조함수(check function)라고 하며 다음과 같이 정의된다.

$$\rho(r) = \frac{1}{2}rI(r\geq 0)+\frac{1}{2}(-r)I(r<0)$$

따라서 대조함수 기반 선형 SV 중위수 회귀분석의 해는 다음과 같은 최적화 문제를 통하여 구해진다.

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

$$\text{subject to} \quad \begin{cases} \frac{1}{2}(y_i - w^t x_i - b) \leq \xi_i \\ \frac{1}{2}(w^t x_i + b - y_i) \leq \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

위의 최적화 문제의 해를 구하기 위해 2절에서와 같이 다음 라그랑즈(Lagrange) 함수

를 최소화한다.

$$L = \frac{1}{2}\parallel w\parallel^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}\alpha_i(\xi_i - \frac{1}{2}y_i + \frac{1}{2}w^t x_i + \frac{1}{2}b)$$

$$-\sum_{i=1}^{n}\alpha_i^*(\xi_i^* + \frac{1}{2}y_i - \frac{1}{2}w^t x_i - \frac{1}{2}b) - \sum_{i=1}^{n}(\eta_i\xi_i + \eta_i^*\xi_i^*)$$

여기서 제약조건 $\alpha_i,\ \alpha_i^*,\ \eta_i,\ \eta_i^* \geq 0$이 성립해야 한다. 따라서 최적화 문제는 다음과 같이 된다.

$$\text{maximize } -\frac{1}{8}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)x_i^t x_j + \frac{1}{2}\sum_{i=1}^{n}y_i(\alpha_i - \alpha_i^*)$$

$$\text{subject to } \frac{1}{2}\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \text{ and } \alpha_i,\ \alpha_i^* \in [0, C].$$

이런 제약조건을 가진 위의 방정식을 푸는 것은 라그랑즈 배수(Lagrange multiplier) $\alpha_i,\ \alpha_i^*$를 구하는 것을 의미한다. 최적인 회귀계수의 추정결과는 다음과 같다.

$$w = \frac{1}{2}\sum_{i=1}^{n}(\alpha_i - \alpha_i^*)x_i, \quad b = -\frac{1}{2}w^t[x_r + x_s]$$

여기서 x_r과 x_s는 스포트 벡터이다. 즉, 추정된 최적의 회귀함수는 다음과 같이 표현된다.

$$f(x) = \frac{1}{2}\sum_{i=1}^{n}(\alpha_i - \alpha_i^*)x_i^t x + b$$

선형의 경우에 해를 구하는 방법을 활용하면 대조함수 기반 비선형 SV 중위수 회귀분석의 해는 다음과 같은 최적화 문제를 통하여 구해진다.

$$\text{maximize } -\frac{1}{8}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(x_i, x_j) + \frac{1}{2}\sum_{i=1}^{n}y_i(\alpha_i - \alpha_i^*)$$

$$\text{subject to } \frac{1}{2}\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \quad \text{and} \quad \alpha_i,\ \alpha_i^* \in [0, C]$$

그리고 최적인 비선형 회귀계수의 추정결과는 다음과 같다.

$$w = \frac{1}{2}\sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\Phi(x_i), \quad b = -\frac{1}{2}\sum_{i=1}^{n}(\alpha_i - \alpha_i^*)[K(x_r, x_i), K(x_s, x_i)]$$

즉, 추정된 최적의 비선형 회귀함수는 다음과 같이 표현된다.

$$f(x) = \frac{1}{2}\sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x_i, x) + b$$

4. 실험결과

본 절에서는 두 종류의 비선형 SV 중위수 회귀모형이 얼마나 회귀함수를 잘 추정하는지에 대한 결과를 시각적으로 보여주기 위해 일변량 비선형 회귀함수 추정문제를 생각한다. 선형 중위수 회귀분석에 대한 결과는 생략한다. 한편, Seok 등 (2002)은 화학공학 관련 한 예제를 통해 SVM이 회귀함수 추정문제에서 신경망 또는 MARS 보

다 더 정확한 추정결과를 제공한다고 설명한다.

위해 10-중 교차타당성(10-fold cross validation) 기법을 사용한다.
본 논문에서는 가우시안 커널을 사용하며, 벌칙상수 C와 커널모수 σ를 결정하기

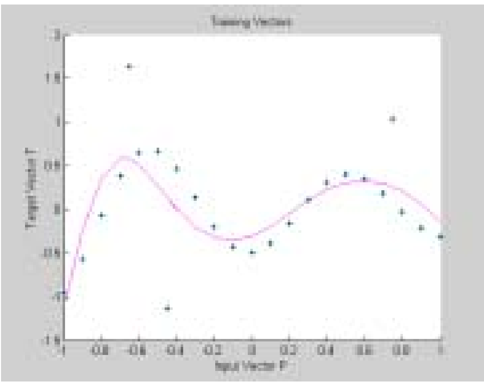

그림 1. 유사 싸인곡선에 대한 신경망 중위수 회귀함수 추정결과

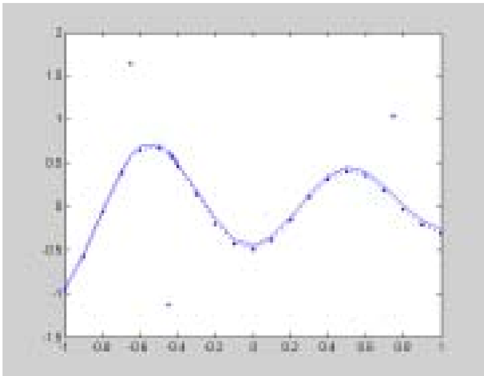

그림 2. 유사 싸인곡선에 대한 SV 중위수 회귀함수 추정결과

예제1. 그림 1과 그림 2는 각각 3개의 이상치를 포함하는 자료에 대해 신경망 및 두 종류의 SV 비선형 중위수 회귀모형이 얼마나 회귀함수를 잘 추정하고 로버스트한지에 대한 결과를 보여준다. 그림에서 ·는 자료점을 나타내며 그림 2의 실선과 점선은 각각 절대편향 오차함수 기반 SV 중위수 회귀함수추정 결과와 대조함수 기반 SV 중위수 회귀함수추정 결과를 보여준다.

그림 1에서 보는 바와 같이 신경망 중위수 회귀모형은 이상치에 의해 많이 영향을 받는 반면에 두 종류의 SV 중위수 회귀모형은 이상치에 거의 영향을 받지 않는 로버스트한 모형임을 알 수 있다. 특히, 대조함수 기반 SV 중위수 회귀모형은 이상치의 영향을 전혀 받지 않는다. 따라서 대조함수 기반 SV 중위수 회귀모형이 절대편향 오차함수 기반 SV 중위수 회귀모형보다 더 안정적인 결과를 주는 것을 확인할 수 있다.

신경망 중위수 회귀모형은 절대편향 오차함수 $|r|$를 사용한다. 그런데 이 오차함수는 점 0에서 미분불능이므로 역전파 알고리즘을 유도하는 것이 불가능하다. 한편 절대편향 오차함수 $|r|$에 대해 다음 관계식

$$|r| = -r\,step(-r) + r\,step(r)$$

이 성립하므로, $step(r)$을 미분가능 함수 $\sigma(r) = 1/(1+e^{-\eta r})$로 근사시키면 역전파 알고리즘을 유도할 수 있다. 본 논문에서는 신경망 중위수 회귀분석을 위한 기존의 복잡한 알고리즘 대신에 전술한 방법을 적용하여 SV 중위수 회귀분석과 비교하였다. 신경망은 5개의 은닉도드를 사용하며, 두종류의 SV 중위수 회귀모형은 10-중 교차타당성 기법에 의해 결정된 모수 $C=500$, $\sigma=0.31$을 사용한다. ■

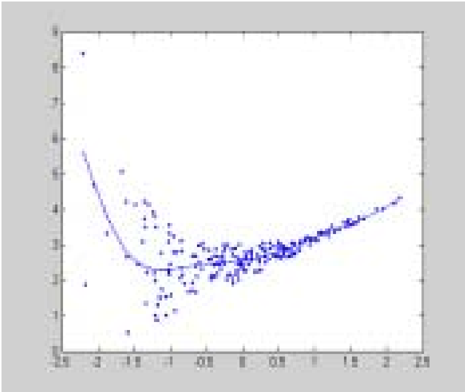

그림 3. $Y = 2 + X + \exp(-X)U$에 대한 SV 중위수 회귀함수 추정결과

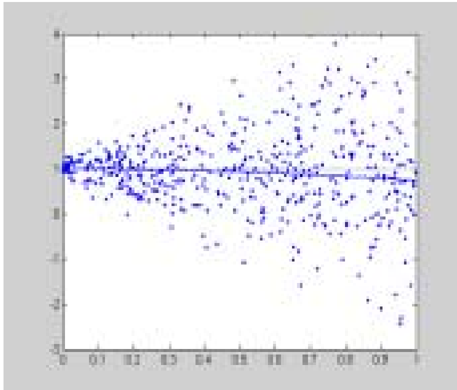

그림 4. $Y \sim N(0, X + X^2)$에 대한 SV 중위수 회귀함수 추정결과

예제2. Yu 등(2001)과 같은 비선형 중위수 회귀분석 관련 참고문헌에 자주 사용되는 다음의 일변량 회귀함수의 추정문제를 생각한다.

$$Y = 2 + X + \exp(-X)U,\ X \sim N(0,1),\ U \sim U(0,1)$$

관측치로 200개의 자료를 생성하여 실험에 사용하였다. 그림 3의 실선과 점선은 각각 절대편향 오차함수 기반 SV 중위수 회귀함수추정 결과와 대조함수 기반 SV 중위수 회귀함수추정 결과를 보여준다. 두 곡선이 거의 일치하는 경향이 있다. 이와 같은 이분산 자료에 대해서 똑같이 좋은 결과를 보여준다.

여기서 두종류의 SV 중위수 회귀모형은 10-중 교차타당성 기법에 의해 결정된 모수 $C = 500$, $\sigma = 1.0$을 사용한다. ■

예제3. 비선형 중위수 회귀분석 관련 참고문헌에 자주 사용되는 다음의 일변량 회귀함수

$$Y \sim N(0, X + X^2),\ X \sim U(0,1)$$

의 추정문제를 생각한다. 500개의 X 자료를 $U(0,1)$로부터 생성하였으며, 그들에 대응되는 500개의 Y 자료는 $N(1, X + X^2)$로부터 생성하였다. 기본적으로 이 자료와 관련된 참 회귀함수는 선형이며 이분산의 성질을 갖는다. 그러나 이런 분포상황을 모른다고 가정하고 SVM 방법을 적용하여 회귀함수를 추정한다. 그 결과는 그림 4에 설명된다.

그림 4의 실선과 점선은 각각 절대편향 오차함수 기반 SV 중위수 회귀함수추정 결과와 대조함수 기반 SVM 중위수 회귀함수추정 결과를 보여준다. 여기서 dashdot 선은 참 회귀직선을 나타낸다. 그림에서 보는바와 같이 두 곡선이 거의 일치하는 경향이 있지만 대조함수 기반 SV 중위수 회귀모형에 의해 추정된 회귀함수가 참 함수에 조금 더 가깝기 때문에 절대편향 오차함수 기반 SV 중위수 회귀모형 보다 좀더 좋은 결과를 보여준다.

두종류의 SV 중위수 회귀모형은 10-중 교차타당성 기법에 의해 결정된 모수

$C=500$, $\sigma=3.0$을 사용한다. ■

결론적으로 3개의 예제를 통해서 볼 때 이분산 자료에 대해서는 절대편향 오차함수 기반 SV 중위수 회귀모형과 대조함수 기반 SV 중위수 회귀모형이 대체로 비슷한 결과를 보여주지만, 이상치가 있는 자료에 대해서는 대조함수 기반 SV 중위수 회귀모형이 좀더 안정된 결과를 보여준다고 말할 수 있겠다. 본 논문에서는 비록 이상치가 있는 자료에 대한 결과만을 보고하였지만 여러 실험을 통해서 볼 때 신경망 중위수 회귀모형에 비해 두종류의 SV 중위수 회귀모형이 대체로 훨씬 더 좋은 결과를 보여주는 것을 확인할 수 있었다. 따라서 회귀함수 추정을 위해 대조함수 기반 SV 중위수 회귀모형의 사용을 추천한다.

참고문헌

1. Gunn, S. (1998), Support Vector Machines for Classification and Regression, ISIS Technical Report, U. of Southampton.
2. Koenker, R. and Bassett G., Regression Quantiles (1978), *Econometrica* 46, 33-50.
3. Koenker R. and Hallock K.F. (2001), Quantile Regression, *Journal of Economic Perspectives*, 40, 4, 129-142, 2001.
4. Seok K., Hwang C. and Cho D. (2002), Prediction Intervals for Support Vector Machine Regression, *Communications in Statistics*, 31, 10, 1887-1898.
6. Smola A. and Schoelkopf B. (1998), On a kernel-based method for pattern recognition, regression, approximation and operator inversion, *Algorithmica*, 22, 211-231.
7. Vapnik V. N. (1995), The Nature of Statistical Learning Theory, Springer, New York.
8. Vapnik V. N. (1998), Statistical Learning Theory, John Wiley & Sons, New York.
9. Yu K., Lu Z. and Stande J. (2001), Quantile regression: applications and current research area (submitted for publication).

[2002년 12월 접수, 2003년 2월 채택]