Proceedings of KFIS Fall Conference, 2003. 5. 3

정해진 기저함수가 포함되는 Nu-SVR 학습방법

Semiparametric Nu-Support Vector Regression

김 영 일, 조 원 희, 박 주 영
고려대학교 제어계측공학과

Young-Il Kim, Won-Hee Cho, Jooyoung Park
Dept. of Control and Instrumentation Engineering, Korea University
E-mail : sstuff@hanmail.net / phone : 011-9048-6356

요 약

ε-SVR(ε-Support Vector Regression)학습방법은 SV(Support Vector)들을 이용하여 함수 근사(Regression)하는 방법으로 최근 주목받고 있는 기법이다. SVM(SV machine)의 한 가지 방법으로, 신경망을 기반으로 한 다른 알고리즘들이 학습과정에서 지역적 최적해로 수렴하는 등의 문제를 한계로 갖는데 반해, 이러한 구조들을 대체할 수 있는 학습방법으로 사용될 수 있다. 일반적인 ε-SVR에서는 학습 데이터와 근사 함수 f사이에 허용 가능한 에러범위 ε값이 학습하기 전에 정해진다. 그러나 Nu-SVR(ν-version SVR)학습방법은 학습의 결과로 최적화 된 ε값을 얻을 수 있다. 정해진 기저함수가 포함되는 ε-SVR 학습방법(Semiparametric SVR)은 정해진 독립 기저함수를 사용하여 함수를 근사하는 방법으로, 일반적인 ε-SVR 학습방법에 비해 우수한 결과를 나타내는 것이 성공적으로 입증된 바 있다. 이에 따라, 본 논문에서는 정해진 기저함수가 포함된 ν-SVR 학습 방법을 제안하고, 이에 대한 수식을 유도하였다. 그리고, 모의 실험을 통하여 제안된 Semiparametric ν-SVR 학습 방법의 적용 가능성을 알아보았다.

Key Words : SVM, ε-SVR, Nu-SVR(ν-SVR), Semiparametric Approach

1. 서론

최근 들어, SVM 학습방법은 지능 시스템 분야에 있어서 중요한 도구로 사용되어 지고 있다. 다양한 분야에 적용되고, 이에 대한 이론이 정립되고 있다. 다른 종류의 지능 시스템을 이용한 학습방법 들과는 달리, 전역 최적해로 수렴할 가능성이 크고, 새로운 입력에 대한 적용이 쉽다. 그리고 학습에 필요한 내부구조가 자동으로 결정된다는 장점이 있다.

Semiparametric Approach(SA)는 Parametric Approach(PA)와 SVR 학습방법이 혼합된 형태로서, SVR을 정해진 기저 함수를 포함하는 형태로 변형시킨 방법이다. 이러한 SA 기법이 적용된 ε-SVR 학습방법은 [3]에서 제안된 바 있다.

본 논문에서는 정해진 기저 함수가 포함된 ν-SVR 학습방법(Semiparametric ν-SVR)에 대하여 제안하고, 그 결과를 일반적인 ν-SVR 학습방법의 경우와 비교하였다.

본 논문의 구성은 다음과 같다. 2장에서는 SVR에 대해 설명하고, 3장에서는 정해진 기저 함수가 포함된 ν-SVR 기법에 대하여 기술한다. 4장에서는 제안된 방법의 유용성을 보이기 위해 모의 실험을 통해 시뮬레이션 및 그 결과를 보인다. 5장에서는 결론에 대해 논의한다.

2. SVR(Support Vector Regression)

2. 1. ε-SVR

주어진 학습 데이터 $\{(x_i, y_i)\}_{i=1}^{m}$에 대해 오차인 $|y_i - f(x_i)|_\varepsilon \triangleq \max\{|y_i - f(x)| - \varepsilon, 0\}$을 최대한 작게 하면서 가능한 한 매끄러운 곡선의 함수 $f(x) = \langle w, x\rangle + b$ 를 구하는 것이 SVR의 목적이다. 일반적으로 이 문제는 최적화(convex optimization)문제로 표현되며, 식으로는 다음과 같이 나타내어진다.[1]

$$\min \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) \tag{1}$$
$$s.t. \quad y_i - (\langle w, x_i\rangle + b) \le \varepsilon + \xi_i$$

$$(\langle w, x_i\rangle + b) - y_i \le \varepsilon + \xi_i^*$$
$$\xi_i\,,\ \xi_i^* \ge 0\,,\ \forall i$$

C는 상수로서 0보다 큰 값을 가지며, 함수 f의 완만함(flatness)과 학습 데이터에 대한 오차값 $\xi_i^{(*)}$ 사이의 trade-off를 결정한다.

이 문제를 해결하기 위해서 식 (1)의 목적 함수와 제약 조건으로부터 라그랑제(Lagrange) 함수를 정의하고 다음과 같이 나타낸다.

$$\begin{aligned} L = {} & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) \\ & - \sum_{i=1}^{m}\alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b) \\ & - \sum_{i=1}^{m}\alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle w, x_i\rangle - b) \\ & - \sum_{i=1}^{m}(\eta_i\xi_i + \eta_i^*\xi_i^*) \\ & (\text{단, } \alpha_i^{(*)}, \eta_i^{(*)} \ge 0) \end{aligned} \tag{2}$$

$w, b, \xi_i^{(*)}$ - 주변수(Primal variable)

$\alpha_i^{(*)}, \eta_i^{(*)}$ - 쌍대변수(Dual variable)

식 (2)의 최적 해는 전체 변수 내에서 안장점이 되므로 [2], 다음 식과 같이 주변수(primal variables)에 대한 부분 미분을 통해 변수를 소거할 수 있다.

$$\frac{\partial L}{\partial w} = 0 \Leftrightarrow w = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)x_i \tag{3}$$
$$\frac{\partial L}{\partial b} = 0 \Leftrightarrow \sum_{i=0}^{m}(\alpha_i - \alpha_i^*) = 0$$
$$\frac{\partial L}{\partial \xi_i^{(*)}} = 0 \Leftrightarrow \alpha_i^{(*)} + \eta_i^{(*)} = C$$

식 (3)의 결과를 식(2)에 대입하여 주변수를 제거하면 다음과 같은 쌍대 문제(Dual Problem)를 얻게 된다.

$$\begin{aligned} \max\ D = {} & -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle \\ & + \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)y_i - \sum_{i=1}^{m}(\alpha_i + \alpha_i^*)\varepsilon \end{aligned} \tag{4}$$
$$s.t.\quad \sum_{i=1}^{m}(\alpha_i - \alpha_i^*) = 0$$
$$\alpha_i^{(*)} \in [0, C]\,,\ \forall i$$

식 (4)의 QP(quadratic programming) 문제를 풀면 최적의 α_i와 α_i^*가 구해지고, 최적의 b값은 Kuhn-Tucker 조건[2]으로부터 얻어진다. 이에 따라, SVR 통하여 얻은 근사 함수 f는 다음과 같이 표현할 수 있게 된다.

$$\begin{aligned} & \therefore f(x) = \langle w, x\rangle + b = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\langle x_i, x\rangle + b \\ & \because w = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)x_i \end{aligned} \tag{5}$$

2. 2. ν-SVR 학습 방법

ε-SVR에서 나타나는 문제점은 ε의 값을 학습하기 전에 정해주어야 한다는 것이다. 그러나 주어진 학습 데이터와 함수 f사이에 오차의 범위를 미리 알 수 없다. 이러한 문제를 해결하기 위해서 ν-SVR에서는 ε의 값을 최적화 문제의 변수로 인식하고, 다른 변수들과 함께 가능한 한 작은 ε의 값을 갖도록 최적화 문제를 표현함으로써, 최적화 문제를 해결함과 동시에 최적의 ε값을 구할 수 있다. 또한, ν값의 변화에 따라 SV의 개수를 조절할 수 있다. 식으로 다음과 같이 나타낸다.

$$\min\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^* + \nu\varepsilon) \tag{6}$$
$$\begin{aligned} s.t.\ & y_i - (\langle w, x_i\rangle + b) \le \varepsilon + \xi_i \\ & (\langle w, x_i\rangle + b) - y_i \le \varepsilon + \xi_i^* \\ & \xi_i^{(*)}\,,\ \varepsilon \ge 0\,,\ \forall i\ \ (\text{단, } \nu \in (0,1]) \end{aligned}$$

식 (6)의 목적 함수와 제약 조건으로부터 라그랑제 함수를 다음과 같이 나타낸다.

$$\begin{aligned} L = {} & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^* + \nu\varepsilon) - \gamma\varepsilon \\ & - \sum_{i=1}^{m}\alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b) \\ & - \sum_{i=1}^{m}\alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle w, x_i\rangle - b) \\ & - \sum_{i=1}^{m}(\eta_i\xi_i + \eta_i^*\xi_i^*) \\ & (\text{단, } \alpha_i^{(*)}, \eta_i^{(*)}, \gamma \ge 0) \end{aligned} \tag{7}$$

식 (7)에서 주변수에 대한 부분 미분을 통해 변수를 소거하면 다음과 같은 쌍대 문제를 얻게 된다.

$$\begin{aligned} \max\ D = {} & -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle \\ & + \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)y_i \end{aligned} \tag{8}$$
$$s.t.\quad \sum_{i=1}^{m}(\alpha_i + \alpha_i^*) \le Cm\nu\,,\quad \sum_{i=1}^{m}(\alpha_i - \alpha_i^*) = 0$$
$$\alpha_i^{(*)} \in [0, C]\,,\ \forall i$$

3. 정해진 기저함수가 포함되는 ν-SVR 학습방법

3. 1. 정해진 기저함수가 포함되는 ε-SVR 학습방법

정해진 기저함수가 포함된 ε-SVR 학습방법은 독립 기저 벡터(independent basis vector)인 $\{\phi_1(\cdot), \ldots, \phi_n(\cdot)\}$를 사용하여 주어진 학습 데이터에 대한 오차를 최대한 작게 하는 함수를 찾는 방법이다. 이것은 다음과 같은 함수 f의 수식으로 나타내어진다.[3]

$$f(x) = \langle w, \psi(x)\rangle + \sum_{i=1}^{m}\beta_i\phi_i(x) \tag{9}$$

함수 f는 SVR에서와 같이 최적화 문제를 해결함으로써 얻어진다.

3. 2. 정해진 기저벡터가 포함되는 ν-SVR 학습방법

식(9)와 [4]에서 소개한 kernels을 갖는 특수한 일반화항을 ν-SVR의 형태로 학습하는 것은, 다음과 같이 최적화 문제로 나타낼 수 있다.

$$\min.\ \ \frac{1}{2}\|w\|^2 + C\sum_{i=m}^{l}(\nu\varepsilon + \xi_i + \xi_i^*) \qquad (10)$$

$$s.t.\ \ \langle w, \phi(x_i)\rangle + \sum_{j=1}^{n}\beta_j\phi_j(x_i) - y_i \le \varepsilon + \xi_i^*$$

$$y_i - \langle w, \phi(x_i)\rangle - \sum_{j=1}^{n}\beta_j\phi_j(x_i) \le \varepsilon + \xi_i$$

$$\xi_i^{(*)}\ ,\ \varepsilon \ge 0\ ,\ \forall i\ (\text{단},\ \nu \in (0,1])$$

최적화 문제를 해결하기 위한 라그랑제 함수는 다음과 같이 나타낼 수 있다

$$L = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m}(\nu\varepsilon + \xi_i + \xi_i^*) + \gamma(-\varepsilon)$$

$$+ \sum_{i=1}^{m}\alpha_i\left\{y_i - \langle w, \phi(x_i)\rangle - \sum_{j=1}^{n}\beta_j\phi_j(x_i) - \varepsilon - \xi_i\right\}$$

$$+ \sum_{i=1}^{m}\alpha_i^*\left\{\langle w, \phi(x_i)\rangle + \sum_{j=1}^{n}\beta_j\phi_j(x_i) - y_i - \varepsilon - \xi_i^*\right\}$$

$$+ \sum_{i=1}^{m}\eta_i(-\xi_i) + \sum_{i=1}^{m}\eta_i^*(-\xi_i^*) = 0 \qquad (11)$$

$$(\text{단},\ \ \alpha_i^{(*)}, \eta_i^{(*)}, \gamma \ge 0\ )$$

위의 식에서 주변수에 대한 부분 미분을 통해 변수를 소거하면 다음과 같은 쌍대 문제를 얻게 된다.

$$\max.\ \ -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)k(x_i, x_j) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*) \qquad (12)$$

$$s.t.\ \ \sum_{i=1}^{m}(\alpha_i + \alpha_i^*) \le Cm\nu\ ,\ \ \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi_j = 0$$

$$\alpha_i, \alpha_i^{(*)} \in [0, C]$$

식(12)에서 $k(x, x')$는 $\langle \phi(x), \phi(x')\rangle$을 나타낸다. 커널 트릭(kernel trick)을 사용하면, 많이 사용되고 있는 가우시안 RBF 커널(radial basis function kernel)에 대하여 다음과 같은 관계가 성립한다.

$$\langle \phi(x), \phi(x')\rangle = K(x, x')$$
$$= \exp(-\|x - x'\|^2/2\sigma^2)$$

식 (12)의 QP(quadratic programming)문제를 풀면 최적의 α_i와 α_i^*가 구해진다. 우리가 구하고자 하는 함수 f는 다음과 같다.

$$f(x) = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)k(x_i, x) + \sum_{j=1}^{n}\beta_j\phi_j(x) \qquad (13)$$

이제 남은 것은 β_i와 ε값을 구하는 것이다. 이것은 KT 조건(Karush-Kuhn-Tucker complementary conditions)에 의해 구할 수 있다.[2]

4. 모의 실험 결과

본 논문에서 제안한 방법의 적용성을 평가하기 위하여, 멕시코 모자(Mexican hat) 함수라고 알려진, $f(x) = sinx + sinc(2\pi(x-5))$에 대하여 함수 근사 문제를 고려해 보았다.

학습 데이터는 멕시코 모자 함수에 noise를 포함하는 $\{(x_i, y_i) \mid y_i = f(x_i) + \xi_i\}$의 형태로 선택하였다. 에러의 표준편차는 0.2로 설정하고, 가우시안 RBF 커널의 표준편차 σ값은 0.25를 사용하였다. trade-off 상수 C값은 0.5로 설정하고, ν값은 0.1에서부터 0.9까지 0.02의 간격으로 나누어 실험하고, 결과를 그림으로 나타내었다. 기저함수로는 $\{\sin x, \cos x, 1\}$을 사용하였다.

그림 2에서는 ν값의 변화에 따라 SV의 개수가 변화하는 결과를 나타내었고, 그림 3과 4에서는 이러한 변화에 따른 오차값을 비교하여 나타내었다. 그리고, 그림 5는 ν값에 변화에 따라 얻어진 ε값의 변화를 나타내었다. 그림 6은 그림 3과 4에서 오차값이 가장 작은 ν값 0.2에 대하여, 일반적인 ν-SVR과 기저 함수과 포함된 ν-SVR 학습방법의 결과를 나타낸 그림이다.

기저함수가 포함된 ν-SVR 학습방법은 일반적인 ν-SVR 방법과 마찬가지로 ν값의 변화에 따라 SV의 개수가 변화하는 특성을 가지면서도, 일반적인 ν-SVR 학습에 비해 오차 값과 ε의 값이 상당히 줄어드는 것을 그림을 통해 확인할 수 있다.

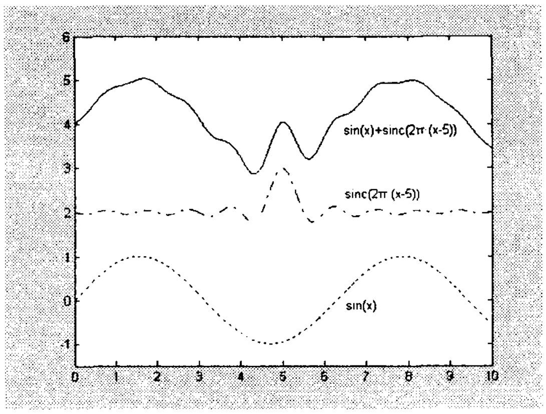

그림 1. '멕시코 모자' 함수

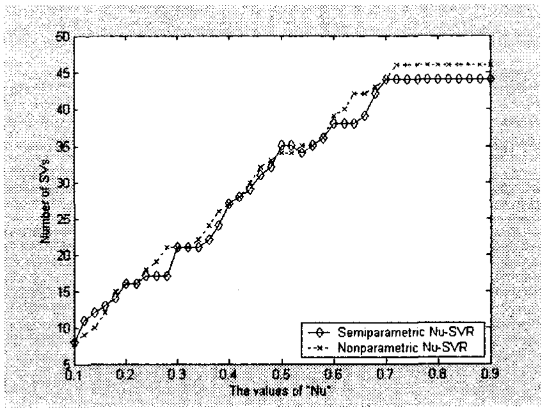

그림 2. Support Vector의 개수 비교

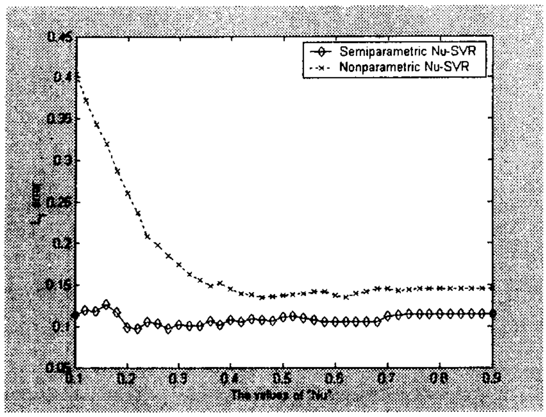

그림 3. L_1 오차값의 비교

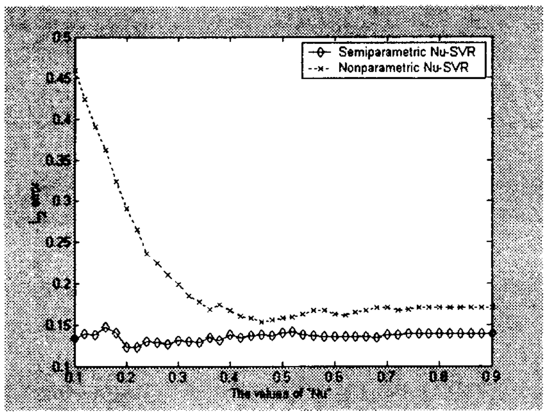

그림 4. L_2 오차값의 비교

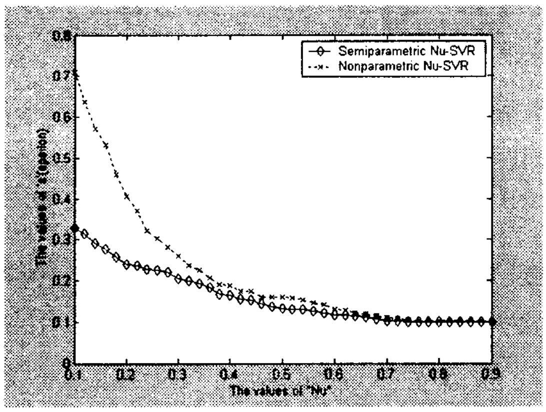

그림 5. 학습의 결과로 얻어진 ε값

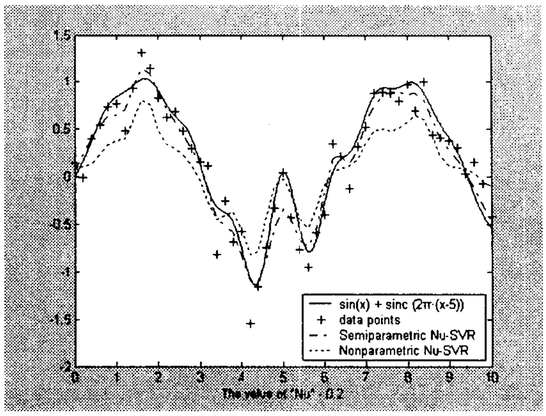

그림 6. v값이 0.2일 때의 학습결과

5. 결론

본 논문에서는 일반적인 ν-SVR를 방법을 이용한 경우와 기저함수가 포함된 ν-SVR방법을 사용했을 경우 각각의 결과를 비교함으로써, SA 기법이 ν-SVR 학습방법에 적용될 수 있다는 것을 제시하고 유도하였다. 제시한 방법론을 간단한 함수 근사 문제에 적용해 본 결과, SA 기법을 이용하여 기존의 ν-SVR 학습방법보다 좋은 결과를 나타내는 것을 확인할 수 있었다. 이와 관련하여 이후에 추가적으로 연구되어져야 할 과제로는, 다양한 예제에 대한 모의 실험을 통하여 장단점을 파악하고, 제안된 방법과 다른 방법들과의 성능을 체계적으로 비교해보는 문제, 그리고 다른 학습방법들에 대해 적용해보는 문제 등을 꼽을 수 있다.

6. 참고문헌

[1] A. Smola and B. Schölkopf, "A tutorial on support vector regression," *Neuro Colt Technical Report NC-TR-1998-030*, Royal Holloway College, University of London, UK, 1998.

[2] N. Cristianini an J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press, 2000.

[3] A. Smola, T. Frieβ, and B. Schölkopf, "Semiparametric Support Vector and Linear Programming Machines," *Nuero COLT Technical Report Series NC2-TR-1998-024*, August, 1998.

[4] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers". In D. Haussler, deitor, *5th Annual ACM Workshop on COLT*, pages 144-152, Pittsburgh, PA, 1992. ACM Press.