회귀용 Support Vector Machine의 효율적인 학습을 위한 조합형 알고리즘

Hybrid Algorithm for Efficient learing of Regression Support Vector Machine

조용현, 박창환†, 박용수†
Yong-Hyun Cho, Chang-Hwan Park†, Yung-Su Park

대구가톨릭대학교 컴퓨터정보통신공학부

요 약

본 논문에서는 SVM의 학습성능 개선을 위해 모멘트와 kernel-adatron 기법이 조합된 하이브리드 학습알고리즘을 제안하였다. 제안된 학습알고리즘은 SVM의 학습기법인 기울기상승법에서 발생하는 최적해로의 수렴에 따른 발진을 억제하여 그 수렴속도를 좀 더 개선시키는 모멘트의 장점과 비선형 특징공간에서의 동작과 구현의 용이성을 가진 kernel-adatron 알고리즘의 장점을 그대로 살리는 것이다. 제안된 알고리즘을 비선형 함수 회귀에 적용해 본 결과 학습속도에 있어서 QP와 기존의 kernel-adatron 알고리즘보다 더 우수한 성능이 있음을 확인하였다.

1. 서론

학습과 대규모 병렬처리에 기인한 신경망을 이용한 접근은 데이터의 overfitting에 따른 모델을 생성할 수 있어 일반화측면에서 어려움이 뒤따르게 된다. 이는 학습을 위한 최적화 알고리즘과 가장 좋은 모델을 선택하는데 이용된 통계적 척도의 결과이다. 이러한 신경망이 가지는 제약들을 해결하기 위해서 support vector machine(SVM)이 제안[2-5]되었으며, 문자, 얼굴, 그리고 물체 인식 등의 실제분야에 성공적으로 적용되었다.

SVM이 overfitting을 효과적으로 막아주는 것은 VC(Vapnik-Chervonenkis) 이론으로 설명될 수 있으며[1], 그것의 학습은 볼록함수를 최대화함으로써 수행되는 것으로 이는 polynomial time내에 발견될 수 있는 유일해가 존재한다는 것을 의미한다. 결국 SVM은 polynomial 학습머신, radial basis 함수 망, 그리고 다층 퍼셉트론(multi-layer perceptron : MLP)[2,3]를 위한 또 다른 학습기법으로 생각될 수 있다.

최근 수많은 실제 분야에서 SVM이 성공적으로 적용되었으나 아직까지 machine learning에서 신경망 등과는 달리 표준도구로 인정받지 못하고 있다. 이는 QP문제를 푸는데 복잡한 계산이 요구되며 시스템을 구현하는데도 어려움이 있기 때문이다. 이러한 제약점을 해결하기 위한 많은 연구가 진행되고 있다[3]. 그 중에서 Campbell 등은 커널(kernel) 방법과 퍼셉트론 규칙을 병합한 반복기법의 kernel-adatron 알고리즘을 제안하였다[3]. kernel-adatron 알고리즘은 adatron의 용이한 구현성과 커널에 의한 비선형 특징공간에서의 동작을 조합한 하이브리드 학습기법이다. 여기서는 특징공간에서의 여유를 최대화하기 위하여 기울기상승(gradient ascent)법을 이용하고 등식의 조건을 만족시키는 라그랑지안 계수(Lagrangian multiplier)을 얻기 위해서 할선법(secant method)을 이용하였다. 그러나 기울기상승법은 해의 변화를 라그랑지안 함수의 미분에 따라 변화시킴으로서 최적해로 수렴될 때 발진이 발생할 수도 있다. 따라서 우수한 구현성은 그대로 살리면서도 최적해로의 수렴에서 발생하는 발진을 막아 좀 더 빠른 속도로 SVM을 학습시킬 수 있는 효과적인 기법의 연구가 요구된다.

본 연구에서는 기울기상승법에서 해의 변화에 과거의 속성을 반영하는 모멘트(momentum) 항을 추가한 수정된 kernel-adatron 알고리즘을 제안하고 이를 SVM의 학습기법으로 이용하였다. radial basis 함수의 커널을 이용한 제안된 학습기법의 SVM을 비선형 함수회귀에 적용하여 시뮬레이션하고 그 타당성을 확인하였으며, QP 알고리즘과 기존의 kernel-adatron 알고리즘 결과와 비교 고찰하였다.

2. Support Vector Machine

SVM은 기존의 신경망 등에서 이용된 경험적 위험을 최소화하는 원리보다는 구조적 위험을 최소화하는 근사적 구현이다. 구조적 위험을 최소화하는 방법은 학습오차 비율의 합으로 범위가 결정되는 시험오차 항과 학습머신의 VC-차원에 의존하는 항에 기반을 두고 있다. 여기서는 회귀문제를 대상으로 소개한다.

데이터 $\{(x_i, y_i), i=1, \cdots, n\} \subset X \times R$가 주어졌다고 가정한다. 여기에서 X는 입력벡터공간 R^d를 나타낸다. 회귀문제는 모든 훈련자료에 대해서 실제 목표값 y_i들로부터 최고 ε만큼의 편차 내에 있으며 가능한 작은 크기의 가중치 벡터 w값을 갖는 다음과 같은 함수 $f(x)$를 찾는 것이다.

$$f(x) = w^T x_i + b \text{ with } w \in X, b \in R \quad (1)$$

윗첨자 T는 벡터의 전치를 나타낸다. 이 식은 가중치 변화에 의해 함수근사를 하는 다층전방향 신경망에서도 활용되고 있다. 이를 위한 한가지 방법은 유클리드 놈(Euclidean norm) $||w||^2$을 최소화하는 것으로 다음과 같은 볼록(convex) 최적화 문제로 간주할 수 있다.

$$\text{minimize} \quad \frac{1}{2}||w||^2,$$

$$\text{subject to} \quad \begin{cases} y_i - w^T x_i - b \le \varepsilon \\ w^T x_i + b - y_i \le \varepsilon \end{cases} \quad (2)$$

여기서 기본가정은 볼록 최적화 문제가 해결 가능하다는 것이나 경우에 따라서는 이 가정이 성립되지 않는다. 따라서 이러한 문제를 해결하기 위해 새로운 슬랙변수 ξ_i와 ξ_i^*를 도입한다. 슬랙변수 ξ_i와 ξ_i^*를 포함한 새로운 최적화 형태는 다음 식 (3)과 같다.

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

$$\text{subject to} \quad \begin{cases} y_i - w^T x_i - b \le \varepsilon + \xi_i \\ w^T x_i + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \quad (3)$$

여기서 C는 양수 값으로 함수 f의 평평함(flatness)과 편차 ε의 허용한계의 균형에 따라 결정된다. 또한 식에서 슬랙변수 ξ_i와 ξ_i^*는 각각 출력에 대한 상한조건과 하한조건을 나타낸다. 위의 식 (3)의 최적화 문제는 다음과 같은 $\varepsilon-$*insensitive* 손실함수(loss function)를 다루는 것에 대응된다.

$$|\xi|_\varepsilon = \begin{cases} 0 & \text{if } |\xi| \le \varepsilon \\ |\xi| - \varepsilon & \textit{otherwise} \end{cases} \quad (4)$$

따라서 식 (3)의 최적화 문제를 좀 더 쉽게 해결하기 위하여 라그랑지안 계수법을 도입하면 식 (3)은 목적 함수(objective function)와 제약조건으로 구성되는 이원문제의 라그랑지안 함수가 된다. 즉,

$$\begin{aligned} L = \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ - \sum_{i=1}^{n}\alpha_i(\varepsilon + \xi_i - y_i + w^T x_i + b) \\ - \sum_{i=1}^{n}\alpha_i^*(\varepsilon + \xi_i^* + y_i - w^T x_i - b) \\ - \sum_{i=1}^{n}(\eta_i \xi_i + \eta_i^* \xi_i^*) \end{aligned} \quad (5)$$

이다. 이때 제약조건 $\alpha_i, \alpha_i^*, \eta_i$, 그리고 η_i^*는 0을 포함한 양수값이어야 한다. 식 (5)를 최소화하는 것은 L의 각 요소 b, w, 그리고 $\xi_i^{(*)}$를 각각 미분하여 얻어질 수 있다. 즉,

$$\partial_b L = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \quad (6)$$

$$\partial_w L = w - \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)x_i = 0 \quad (7)$$

$$\partial_{\xi_i^{(*)}} L = C - \alpha_i^{(*)} - \eta_i^{(*)} = 0 \quad (8)$$

이다. 얻어진 식 (6), (7), (8)을 식 (5)에 대입하여 α_i와 α_i^*를 각각 구할 수 있다. 결국 식 (5)를 정리하면

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

$$-\varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n}y_i(\alpha_i - \alpha_i^*) \quad (9)$$

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

이다. 여기서 제약조건을 만족하는 방정식을 푸는 것은 식 (9)를 최대화하는 라그랑지안 계수 α_i와 α_i^*를 구하는 것이다.

이상에서 살펴 본 회귀함수는 선형의 경우에만 국한하여 설명한 것으로 모든 문제에 적용될 수는 없다. 그러므로 비선형 회귀함수에 적용하기 위해서는 좀 더 일반적인 형태의 SVM이 고려되어야 한다. 이를 위한 방법으로 SVM에서 입력 x를 고차원의 특징공간 z로의 사상을 이용한다. 그러나 이 방법은 고차원 특징공간에서의 내적 $(z(x)\cdot z(x_i))$의 계산이 요구된다. 어떤 조건 하에서, 내적 계산은 아주 비효율적이지만 커널함수 K를 사용함으로써 효율적인 계산이 가능하다. 여기서 커널함수와 특징공간과의 관계는 다음 식 (10)과 같다.

$$K(x, x_i) = z(x)^T \cdot z(x_i) \quad (10)$$

사상을 위한 커널함수로는 polynomial 함수, RBF 함수, 그리고 s-자형 함수 등이 이용된다.

한편, $\varepsilon-$*insensitive* 손실함수(loss function)를 사용하는 SVM의 비선형 회귀함수의 해는 다음과 같이 구해진다. 즉,

$$\begin{aligned} \text{maximize} \quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(x_i, x_j) \\ -\varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n}y_i(\alpha_i - \alpha_i^*) \end{aligned} \quad (11)$$

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

이다. 이는 선형 회귀함수의 표현식 (9)에서 입력벡터의 내적을 커널함수로 대치한 것에 불과하다. 여기에서도 방정식을 푸는 것은 식 (11)을 최대화하는 라그랑지안 계수 α_i와 α_i^*를 구하는 것으로 선형적인 회귀함수의 경우와 동일하다. 또한 특징공간에서 최적의 회귀함수식은 다음과 같이 나타낼 수 있다. 즉,

$$f(x) = \sum_{SVs}(\overline{\alpha_i} - \overline{\alpha_i^*})K(x_i, x) + \overline{b} \quad (12)$$

이다.

결국 구해진 라그랑지안 최적의 계수 $\overline{\alpha_i}$와 $\overline{\alpha_i^*}$를 이용한 최적의 가중치 벡터 $\overline{w}$와 $\overline{b}$는 각각

$$\overline{w} = \sum_{SVs}(\overline{\alpha_i} - \overline{\alpha_i^*})z(x_i),$$

$$\overline{b} = -\frac{1}{2}\sum_{SVs}(\overline{\alpha_i} - \overline{\alpha_i^*})[K(x_r, x_i), K(x_s, x_i)] \quad (13)$$

로 계산될 수 있다.

지금까지 살펴본 SVM을 이용한 최적의 회귀함수를 구하는 문제는 최적화 패키지를 사용하여 해결할 수 있는 QP 문제로 변환하여 해결하는 것이다. 그러나 QP 과정에서는 계산적인 복잡성과 구현의 어려움이 있으며, 문제의 규모가 커질수록 그 제약점은 더욱 더 심화된다. 결국 좀 더 빠르고 구현이 용이한 새로운 해결책이 요구된다.

3. 모멘트와 Kenel-Adatron이 조합된 학습알고리즘

Campbell 등이 제안한 kernel-adatron 알고리즘은 퍼셉트론과 유사한 알고리즘인 adatron과 비선형 특징공간에서의 동작을 위한 kernel을 조합한 하이브리드 학습기법이다. 하지만 kernel-adatron알고리즘에서는 라그랑지안 함수

$$L(\alpha) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(x_i, x_j) - \varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*) \quad (14)$$

의 최대화를 위한 α와 α^*의 변화량 $\delta\alpha_k$와 $\delta\alpha_k^*$의 계산에 기울기상승법을 그대로 이용하고 있어 최적해로 수렴할 때 학습률 η에 따라서는 발진이 발생할 수도 있다. 따라서 수렴에 따른 발진을 막아 줄 수만 있다면 좀 더 빠르게 최적해로 수렴시킬 수 있을 것이다. 이를 위해 본 연구에서는 $\delta\alpha_k$와 $\delta\alpha_k^*$를 계산할 때 이전 변화의 속성을 반영하는 모멘트를 추가하였다. 이는 현재의 계산방향이 이전의 계산방향을 따르도록 하는 것이다. 이렇게 하면 kernel-adatron 알고리즘이 가지는 우수한 구현성은 그대로 살리면서도 최적해로의 수렴에서 발생하는 발진을 억제시켜 좀 더 빠른 속도로 SVM을 학습시킬 수 있다. 제안된 모멘트를 이용한 $\delta\alpha_k$와 $\delta\alpha_k^*$의 계산식은 각각

$$\delta\alpha_k(t+1) = \eta_1\frac{\partial L}{\partial\alpha_k(t)} = \eta_1\left(-\sum_{i=1}^{n}(\alpha_i(t) - \alpha_i^*(t))K(x_k, x_i) - \varepsilon + y_k\right) + m(\alpha_k(t) - \alpha_k(t-1)) \quad (15)$$

$$\delta\alpha_k^*(t+1) = \eta_2\frac{\partial L}{\partial\alpha_k^*(t)} = \eta_2\left(-\sum_{i=1}^{n}(\alpha_i(t) - \alpha_i^*(t))K(x_k, x_i) - \varepsilon - y_k\right) + m(\alpha_k^*(t) - \alpha_k^*(t-1))$$

와 같으며, 여기서의 η와 m은 각각 학습률과 모멘트이다.

4. 시뮬레이션 결과 및 분석

제안된 알고리즘을 이용한 회귀용 SVM의 성능을 평가하기 위해 실험에서는 커널함수로 가우스형태의 radial basis 함수 $K(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right)$를 이용하였다.

실험은 -10에서 10의 구간에서 51개의 등간격 데이터를 추출한 비선형 함수 $f(x) = \sin c(x)$를 대상으로 하였으며, 기존의 표준 QP 알고리즘 및 kernel-adatron 알고리즘과의 결과와 비교 고찰하였다.

그림 1은 $\varepsilon = 0$, $\eta = 0.5$, $m = 0.01$로 하여 radial basis 함수의 폭 σ의 변화에 따른 실제 데이터 값과 근사화된 값과의 절대합오차(Absolute Sum Error : ASE) $= \sum_{i=1}^{n} | y_i - f(x) |$ 를 나타낸 것이다. 그림에서 보면 QP 알고리즘은 σ값에 무관하게 일정한 절대합오차값을 나타내나 kernel adatron알고리즘과 제안된 알고리즘에서 성능은 상대적으로 QP 알고리즘보다 커널폭 σ에 더욱 의존됨을 알 수 있다. 그림에서 제안된 알고리즘과 kernel-adatron 알고리즘의 경우, σ의 값은 0.3에서 0.6 사이가 가장 적당함을 알 수가 있다.

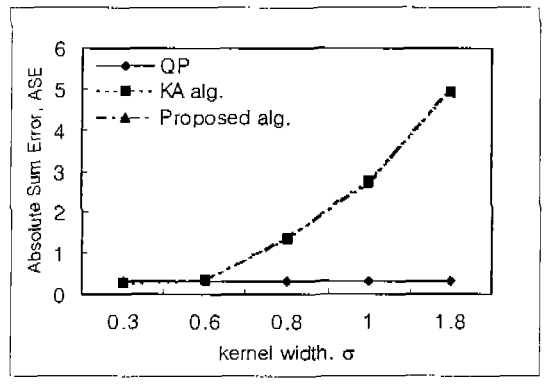

그림 1 radial basis 함수의 폭 σ의 변화에 따른 SVM의 오차합

그림 2는 $\varepsilon = 0$, $\eta = 0.5$, $\sigma = 0.6$, $m = 0.01$로 하여 대상함수에 대한 QP, kernel-adatron 알고리즘, 그리고 제안된 알고리즘의 회귀결과를 보여 주는 그래프이다. kernel-adatron 알고리즘과 제안된 알고리즘은 거의 유사하게 근사화된다는 것을 나타낸다. 하지만 QP는 kernel-adatron 알고리즘과 제안된 알고리즘보다 상대적으로 근사화가 떨어진다는 것을 알 수가 있다.

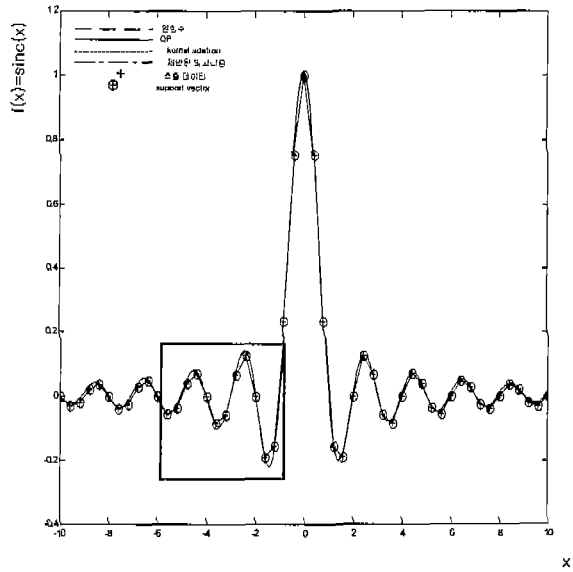

그림 2 $f(x) = \sin c(x)$에 대한 QP, KA 그리고 제안된 알고리즘의 회귀함수

그림 2의 사각부분을 확대한 그림 3을 보면 더 정확하게 알 수가 있는데, 특히 함수의 변화가 심한 함수의 양쪽 끝부분에서 근사를 살펴보면 kernel-adatron 알고리즘과 제안된 알고리즘은 비교적 원함수와 거의 유사하게 근사화되는 것에 비해서, QP의 경우는 추출된 데이터에 선

형적으로 근사되는 것을 볼 수 있다.

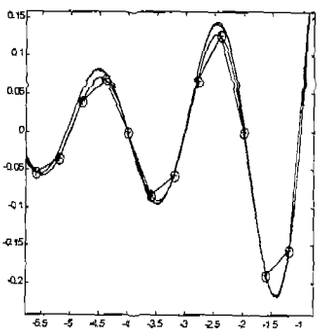

그림 3 그림 2 사각 영역의 확대

즉, 추출데이터의 수가 많은 구간에서는 QP보다 kernerl-adatron 알고리즘과 제안된 알고리즘이 더 우수한 성능을 보여준다는 것을 알 수 있다. 또한 이를 좀 더 구체적으로 비교하기 위해서 학습시간과 ASE를 각각 측정한 결과, QP 알고리즘, kernel-adatron 알고리즘, 제안된 알고리즘의 ASE는 각각 0.31, 0.36, 0.37이고, 학습 시간은 QP 알고리즘 t_Q=4.9, kernel-adatron 알고리즘 t_k=0.16, 제안된 알고리즘 t_p=0.06초가 걸렸다. 이는 QP 알고리즘에 비해서 kernel-adatron 알고리즘과 제안된 알고리즘이 빠른 학습도를 가짐을 알 수 있다. 제안된 알고리즘은 keren-adatron 알고리즘에 대해 약 2.7배 정도, QP알고리즘에 대해서는 약 81.7 배 정도 개선되었음을 알 수 있다. 따라서 kernel-adatron 알고리즘과 제안된 알고리즘만을 비교하면 절대합오차로 표현되는 회귀성능은 유사하나 학습속도에서는 훨씬 개선된 성능이 있음을 알 수 있다.

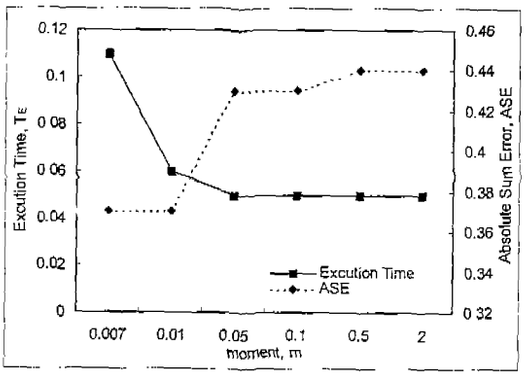

그림 4 모멘트의 변화에 따른 제안된 알고리즘의 학습시간과 ASE

그림 4는 모멘트 m의 변화에 따른 제안된 알고리즘의 학습시간 T_E와 절대합오차 ASE를 나타낸 것이다. 그림에서 보면 m이 0.01일 때 학습시간과 ASE에서 가장 우수한 성능을 보임을 알 수 있다. 하지만 다른 모멘트 값에서도 학습시간과 ASE측면에서 제안된 알고리즘은 기존의 kernel-adatron 알고리즘보다 우수한 성능이 있음을 알 수 있다. 이는 동일 문제에 대해서 kernel-adaron 알고리즘의 경우 학습시간이 0.16인데 비해 그림에서는 최대 0.11 정도로 0.05 정도의 빠른 학습시간 차를 나타내기 때문이다.

이상에서 제안된 방법은 기존의 QP 알고리즘이 가지는 학습속도와 구현의 어려움 및 기존 kernel-adatron 알고리즘에서 발생되는 발진을 동시에 해결할 수 있음을 확인할 수 있다.

5. 결론

본 연구에서는 회귀용 SVM의 효율적인 학습을 위해 기울기상승법에서 해의 변화에 과거의 속성을 반영하는 모멘트 항을 추가한 조합형 알고리즘을 제안하였다. 이는 기울기상승법에서 발생하는 최적해로의 수렴에 따른 발진을 억제하여 그 수렴속도를 좀 더 개선시키는 모멘트의 장점과 kernel-adatron 알고리즘의 구현 용이성을 그대로 살리기 위함이다.

제안된 학습알고리즘의 SVM을 비선형 회귀함수에 적용하여 시뮬레이션 한 결과, QP 알고리즘과 기존의 kernel-adatron 알고리즘보다 회귀성능과 학습시간에 있어서 우수한 결과를 보였다.

앞으로 제안된 알고리즘을 좀더 큰 규모의 문제에 적용하는 것과 다양한 응용분야에 그 성능을 확인해 보아야 할 것이다.

참고문헌

[1] V. Vanpnik, The Nature of Statistical Learning Theory, Springer Vergag, 1995

[2] M.O. Stitson, J.A.E. Weston, A. Gammerman, V. Vovk, and V. Vapnik," Theory of Support Vector Machines," Technical report CSD-TR-96-17, Royal Holloway, Univ. of London, May1998

[3] A.J.Smola,and B.Scholkofp, "A Tutorial on Support Vector Regression," NeuroCOLT2 Technical Report, NeuroCOLT, 1998.

[4] C.Campbell and N.Cristianini, "Simple Learning Algorithms for Training Support Vector Machines," Dept. of Engineering Mathematics Technical Trport, U.of Bristol, 1998.

[5] S.Gunn, "Support Vector Machines for Classification and Regression," ISIS Technical Report, U.of Southampton, 1998.

[6] Thiol-Thomas Frieb, N.Cristianini and C.Campbell "The Kernel-Algorithm: a Fast and Simple Learning Procedure for Support Vector Machines,"

[7] N.Cristianini, C.Campbell and J.Shawe-Taylor "Dynamically Adapting Kernels in Support Vector Machines," Produced as part of the ESPRIT Working Group in Neural and Computational Learning II, NeuroCOLT2 27150.