Journal of the Korean
Data & Information Science Society
2001, Vol. 12, No. 2, pp. 95 ~ 102

서포터벡터학습의 효율적 알고리즘 [1]

석경하 [2]

요 약

최적의 SVM 가중치를 선택하는 방법 중에서 메모리와 속도의 문제를 해결하는 방법 중 하나가 커널애더트론 방법(Kernel Adatron, KA)이다. 본 연구에서는 KA방법을 제곱무감각손실함수까지 확장을 한 알고리즘을 개발한다. 그리고 추정해야 될 라그랑제 배수(Lagrange multiplier)의 수를 반으로 줄이는 알고리즘을 제시한다. 그리고 제시된 알고리즘의 효율성을 여러 모의실험을 통해서 입증한다.

1. 서론

신경망을 이용한 문제 해결방법은 아주 폭넓게 이용되고 있는 현실이다. 그러나 신경망을 학습시키는 과정에서 과대적합(overfitting)을 따르는 모형을 생성할 수 있어 일반화에 어려움이 많은 실정이고[5], 또한 학습에 중요한 역할을 하는 모수(parameter)들을 설정하는 과정이 객관적이고 분석적인 방법을 거치는 것이 아니라 사용자의 경험에 의존하는 문제점을 가진다. 그러므로 문제에 따라 사용자에 따라 해석이 달라지는 문제점을 가진다. 그 뿐만 아니라 국소최적점에 빠지는 학습으로 인해 최적의 해를 구하지 못하는 예가 종종 있다. 이러한 신경망이 가진 단점을 보완하기 위하여 개발된 Support Vector Machine(SVM)은 VC이론을 근거로 하여 개발되었다[5,6,10,11]. 그로 인해 예측을 잘 하고 또한 벌칙항(penalty)을 이용하여 과대적합(overfitting)을 피하는데 성공적이라는 것이 여러 연구의 결과에서 보여주고 있다. 그리고 문자인식, 얼굴인식등 많은 응용분야에서도 많은 성공사례를 보여주고 있다[1,5,9]. SVM이 신경망과 달리 과대적합을 피할 수 있는 것은 VC이론으로 설명이 될 수 있으며, 또한 SVM은 볼록함수(convex function)를 최소화하는 학습을 진행하기 때문에 신경망과는 달리 유일한 최적의 해를 구할 수 있다는 장점을 가지고 있다. 또한 함수근사의 문제에 있어서 이상치에 둔감하다는 장점도 가지고 있다.

여러 응용분야에서 많은 성공사례를 보여주고 있음에도 불구하고 우리가 원하는 해를 구하기 위해서 사용해야하는 QP(Quadratic Programming)문제를 해결하는데 많은 시간과 메모리가 요구되기 때문에 기계학습의 표준도구로 이용되는데 어려움이 있는 실정이다. 이 문제를 해결한 알고리즘 중 하나가 커널애더터론(Kernel Adatron, KA) 방법[2,3]인데 손실함수를 선형 ε-무감각 손실함수(Linear ε-*insensitive* loss function, LIL)를 사용하였다. 그리고 두 종류의 라그랑제 배수를 따로 최신화시키는 알고리즘을 사용했다.

이 논문에서는 LKA(LIL에 의한 KA알고리즘)을 제곱 ε-무감각손실함수(Quadratic ε-*insensitive* loss function, QIL)로 확장시키고, 계산시간을 줄이기 위해 라그랑제 배수를

[1]본 논문은 2000년도 인제대학교 해외연수비 지원에 의한 것임
[2]경남 김해시 어방동 인제대학교 데이터정보학과

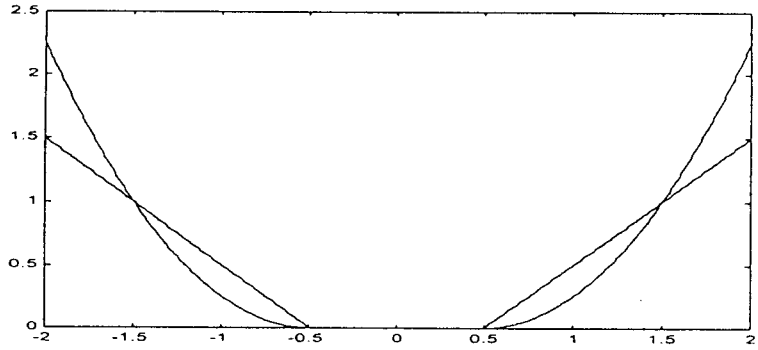

그림 1. 선형 ε-무감각 손실함수와 제곱 ε-무감각 손실함수

동시에최신화시키는 알고리즘, QKA(QIL에 의한 KA 알고리즘)을 제시한다. 그리고 제시된 알고리즘의 효율성을 모의실험을 통해서 입증한다. 위의 그림 1에서 선형 ε-무감각 손실함수와 제곱 ε-무감각손실함수의 형태를 보여준다.

2. 서포터벡트학습을 이용한 함수근사

SVM은 원래 분류를 위해 Vapnik과 공동연구자들에 의해 개발되었으며 많은 응용분야에서 좋은 결과를 보여주고 있어 그 이용이 점점 더 확대되고 있는 실정이다. SVM에는 분류를 위한 것과 함수근사[4]를 위한 두 종류가 있는데 분류를 위한 것은 함수근사의 특별한 것으로 간주되기 때문에 본 연구에서는 함수근사에 관한 것만 연구하기로 한다.
훈련자료 $\{(x_i, y_i), i = 1, ..., n\} \subset \mathbf{X} \times \mathbf{R}$ 주어졌다고가정한다. 여기서 $\mathbf{X}$는d차원 입력벡터 공간 $\mathbf{R}^d$를 나타낸다. 모든 훈련자료에 대해서 실제목표값 y_i 들로부터 최고 ε(무감각모수, insensitivity parameter)만큼의 편차내에 있으며 가능한 작은 크기의$\mathbf{w}$값을 갖는 다음과 같은 함수

$$f(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + b \; with \; \mathbf{x} \in \mathbf{X}, b \in \mathbf{X}$$

가 SVM이다. 이 이론은 가중치 감소(weight decay)를 사용하여 함수근사를 하는 다층신경망에서도 활용되고 있다. 이를 해결하기 위한 한가지 방법은 유클리드 노름(Euclidian norm)의 제곱 $\|\mathbf{w}\|^2$을 최소화시키는 것이다. 이러한 문제는 공식적으로 다음과 같은 볼록최적화(convex optimization)문제로 간주 할 수 있다[5,8,9].

$$minimize \; \frac{1}{2}\|\mathbf{w}\|^2,$$

$$subject \; to |y_i - \mathbf{w}^t\mathbf{x_i} - b| \leq \varepsilon$$

여기서 기본 가정은 볼록최적화 문제의 해결이 가능하다는 것이다. 그러나 가끔 이 가정이 성립되지 않는 경우가 있다. 따라서 이를 해결하기 위해서 새로운 slack변수 ξ, ξ^*를도입하여 Vapnik[10,11]이 제안한 최적화문제를 QIL에 적용을 시키면 다음과 같은 최적화 문제

가 된다.

$$\begin{aligned} &\textit{minimize} && \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\xi_i^2 + \xi_i^{*2}), \\ &\textit{subject to} && y_i - \mathbf{w}^t\mathbf{x_i} - b \le \varepsilon + \xi_i,\ i =, \ldots, n \\ &&& \mathbf{w}^t\mathbf{x_i} + b - y_i \le \varepsilon + \xi_i^*,\ i =, \ldots, n \\ &&& \xi_i, \xi_i^* \le 0, \qquad\qquad i =, \ldots, n \end{aligned} \tag{1}$$

여기에서 C(¿0)는 f의 평평한 정도와 추정오차의 크기를 조절하는 일반화모수(generalization parameter, regularization parameter)이다. (1)식에서 ξ_i^2, ξ_1^{*2} 대신에 ξ, ξ^*사용하면 LKA가 된다. 이 문제를 라그랑제함수(Lagrangian function)로 표현을 하면

$$\begin{aligned} \mathrm{L} \;=\; & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\xi_i^2 + \xi_i^{*2}) - \sum_{i=1}^{n}\alpha_i(\varepsilon + \xi_i - y_i + \mathbf{w}^t\mathbf{x}_i + b) \\ & - \sum_{i=1}^{n}\alpha_i^*(\varepsilon + \xi_i^* + y_i - \mathbf{w}^t\mathbf{x}_i - b) - \sum_{i=1}^{n}(\eta_i\xi_i + \eta_i^*\xi_i^*) \end{aligned} \tag{2}$$

이다. 여기서 제약조건은 $\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \ge 0$ 이다. (2)식의 라그랑제 함수를 최소화하는$\mathbf{w}$와 b를 구하여 (2)식에 다시 대입을 하고 커널을 이용한 비선형함수의 추정으로 확장을 하면 다음의 최대화문제가 된다.

$$\begin{aligned} &\textit{maximaize} && -c\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\{K(\mathbf{x}_i, \mathbf{x}_j + \delta_{ij}/C)\} \\ &&& -\varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n}y_i(\alpha_i - \alpha_i^*), \\ &\textit{subject to} && \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \textit{ and } \alpha_i,\ \alpha_i^* \ge 0. \end{aligned} \tag{3}$$

많이 사용되고 있는 커널함수는

$$K_p(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^t\mathbf{y} + 1)^p,\ K_\sigma(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2/\sigma^2)$$

인데, 여기서 p와 σ가 커널평활량(smoothing parameter)으로서 SVM에서 중요한 역할을 한다. 이런 제약 조건 하에서 위의 방정식을 푸는 것은 라그랑제 배수 α_i, α_i^*를 구하는 것이며 최적의 함수 근사는 아래와 같다.

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

여기에서 $\bar{\alpha_i}$와$\bar{\alpha_i^*}$는 (3)식의 해이고 $\bar{b}$는 $\alpha_i - \alpha_i^* \ne 0$ 인 i에 대해 $f(\mathbf{x}_i) - y_i = -\varepsilon - (\alpha_i - \alpha_i^*)/C$를 만족하는 값이다.

SVM을 사용한 최적의 함수근사 문제는 QP문제로 귀결이 되는데 이는 많은 메모리와 계

산시간을 요구한다[2,3]. 이러한 QP의 단점을 해결하기 위해 개발된 방법이 LKA방법인데 LIL에 기초를 두고 있고 번째 관측치에 대응되는 라그랑제 배수 α_i,α_i^*를

$$\alpha_i = \alpha_i + \delta\alpha_i,\ \alpha_i^* = \alpha_i^* + \delta\alpha_i^* \tag{4}$$

을 반복하여 구한다. 그러나 본 논문에서 제안하고자 하는 QKA방법은 Karush-Kuhn- Tucker 조건

$$\alpha_i\alpha_i^* = 0,\ \ i = 1, \cdots, n$$

과 $\beta_i = \alpha_i - \alpha_i^*$을 이용하여 다음과 같은 최대화문제로 귀착을 시킨다.

$$\begin{aligned} &\textit{maximaize} \quad L(\beta) = -\frac{1}{2}\sum_{i,j=1}^{n}\beta_i\beta_j\{K_\sigma(\mathbf{x}_i\mathbf{x}_j) + \delta_{ij}/C\} - \varepsilon\sum_{i=1}^{n}|\beta_i + \sum_{i=1}^{n}y_i\beta_i \\ &\textit{subject to} \quad \sum_{i=1}^{n}\beta_i = 0. \end{aligned} \tag{5}$$

이렇게 고친 이유는 (3)식의 과정을 거치면 α_i와α_i^*두 개의 라그랑제 배수를 최신화 해야 하지만 (5)식에서는 하나의 라그랑제 배수만 최신화 하면 되기 때문이다. 이 때 라그랑제 배수 β_k의 변화량을 구하면

$$\delta\beta_k = \eta\{y_k - sign(\beta_k)\varepsilon - \sum_{i=1}^{n}\beta i[K_\sigma(x_i, x_k) + \delta_{ik}/C]\},$$

가 된다.여기에서 $-sign(x) = 1,\ x \geq,\ = -1, x \leq,\ = 0,\ otherwise.$ 당연히 하나의 값만 최신화 시키는 과정이 계산시간을 줄일 수 있겠다. 그리고 최신화시키는 과정에서 학습률 η때문에 일어날 수 있는 진동을 막아주고 좀 더 빠른 학습을 위해서 모멘텀(momentum)항을 추가한 아래와 같은 변화량을 고려한다.

$$\begin{aligned} \delta\beta_k(t+1) &= \eta_1\{y_k - sign(\beta_k(t))\varepsilon - \sum_{i=1}^{n}\beta i(t)[K_\sigma(x_i, x_k) + \delta_{ik}/C]\} \\ &\quad + \eta_2\{\beta_k(t) - \beta_k(t-1)\} \\ &= \eta_1\{y_k - \hat{y_k}(t) - sign(\beta_k(t))\varepsilon - \beta_k(t)/C\} \\ &\quad + \eta_2\{\beta_k(t) - \beta_k(t-1)\} \end{aligned} \tag{6}$$

여기에서 t는 반복수(epoch)이며 η_1과η_2는 학습률로 양의 값이다.
다음절에서는 모의 실험을 통하여 (6)식을 이용한 SVM의 학습알고리즘이 효율적이라는 것을 보여 주도록 한다.

3. 모의실험

제안된 알고리즘의 성능을 분석하기 위해 비선형 일변량 함수의 근사문제를 생각한다. 근사를 위해 두 개의 함수를 예로 들어 설명하기로 한다. 첫 번째 함수는 Wahba & Wold[12]의 French 곡선 $f(x) = 4.26(e^{-x} - 4e^{-2x} + 3e^{-3x})$이다. 입력값으로는 구간 [0,3]에서 100개의 점을 등간격으로 추출하여 사용하였다. 두 번째 함수는 Shao[7]의 논문에서 사용된 다음의 불연속 다항함수이다.

$$f(x) = \begin{cases} 4x^2(3-4x), & x \in [0, 0.5], \\ \frac{4}{3}x(4x^2-10x+7) - \frac{2}{3}, & x \in [005, 0.75], \\ \frac{16}{3}x(x-1)^2, & x \in [0.75, 1] \end{cases}$$

입력값으로는 구간 [0,1]에서 100개의 점을 등간격으로 추출하였다. 모의실험에서 사용된 커널은 $K_\sigma(\mathbf{x},\mathbf{y}) = \exp(-\|\mathbf{x}-\mathbf{y}\|^2/\sigma^2)$이다. 함수근사를 위한 SVM에서 중요한 모수인 C,ε,σ는 VC-이론을 근거로 교차타당성방법(cross-validation)을 사용하여 선택하였는데[8,11] 첫 번째함수의 경우에는 C=50, $\varepsilon = 0.001$, $\sigma = 0.2$ 가 적절하였고 두 번째는 가 C=10, $\varepsilon = 0.001$, $\sigma = 0.05$ 적절하였다. 그리고 학습율은 $\eta_1 = 0.02$, $\eta_2 = 0.01$ 을 사용하였다.
그림 2는 이상치가 없는 경우의 French 곡선에 대해 QPSVM(QP를 사용한 SVM)과 제안된 알고리즘을 사용한 SVM(QKASVM)의 함수근사의 결과를 보여준다. 모든 아래의 그림에서와 마찬가지로 +는 자료점을 나타내고 짧은 절단선과 긴 절단선은 각각 QPSVM과 100회반복의 QKASVM에 의한 함수근사의 결과를 보여준다. 세 종류의 선은 거의 일치하여 구별할 수 없을 정도로 차이가 없는 좋은 결과를 보여준다.

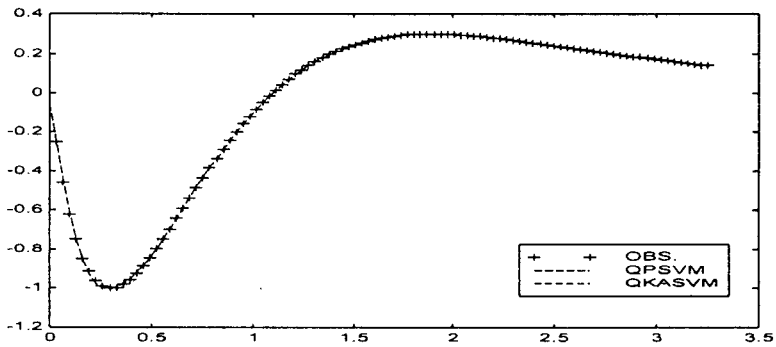

그림 2 이상치가 없는 French 곡선의 QPSVM과 QKASVM

그림3은 이상치가 있는 경우의 French 곡선에 QPSVM과 QKASVM의 함수근사 결과를 보여준다. 여기에서 주목 할 사실은 QPSVM은 이상치에 영향을 받지 않은 반면 QKASVM

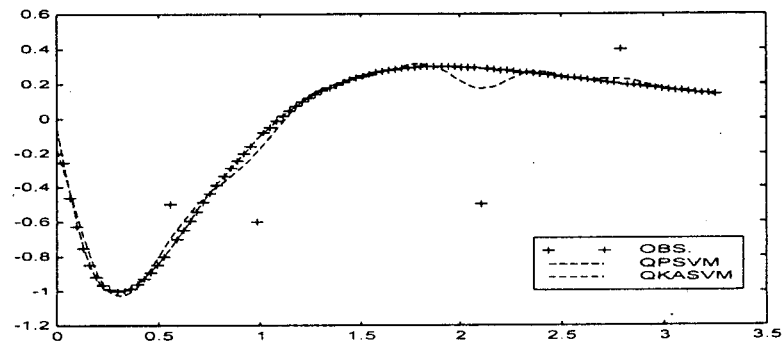

그림 3 이상치가 있는 French 곡선의 QPSVM과 QKASVM

의한 근사는 이상치에 영향을 받고 있다는 것이다. 이는 적합도 함수로 사용된 QIL가 이상치에 영향을 받을 수 있는 형태라는 점을 인식 할 때 당연한 결과라고 볼 수 있다. QPSVM에 사용된 손실함수는 선형이므로 로버스트한 형태의 손실함수 일 뿐 아니라 LKA에서 유도된 라그랑제 배수가 가지는 성질 때문에 QPSVM은 로버스트한 결과를 보여준다. 다음으로는 이상치가 없는 경우의 불연속 다항함수의 근사를 그림4에서 보여준다. 여기에서도 위에서와 마찬가지로 QPSVM과 QKASVM에 의한 근사치가 0.5근방을 제외한 영역에서 매우 유사한 형태를 나타낸다. 대체적으로 불연속점인 0.5근방을 제외한 대부분의 영역에서 좋은 결과를 보여주지만 0.5근방에서 QPSVM이 약간 우수함을 알 수 있다.

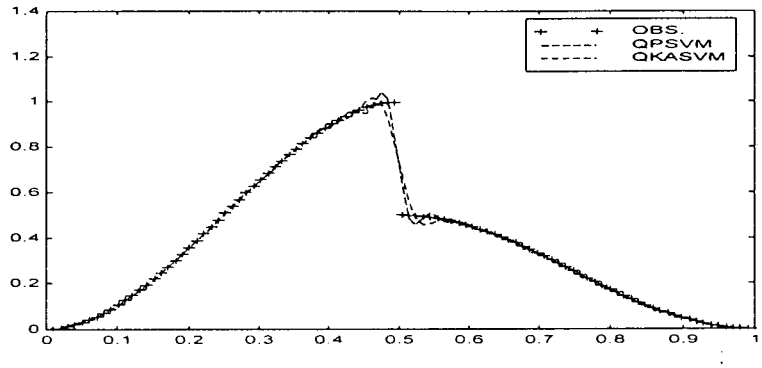

그림 4. 불연속 다항곡선의 QPSVM과 QKASVM

함수근사의 경우에는 커널평활량 σ의 값을 작게 함으로써 근사의 정확도를 높일 수 있다. 그러므로 잡음(오차)이 있는 경우의 함수추정에서 성능비교를 통하여 두 알고리즘을 비교하도록 한다.

다음의 그림은 잡음이 있는 French곡선의 추정을 보여준다. 그림에서 실선은 French 곡선이고 자료점은 +로 표시되었는데 French곡선에 $N(0, 0.2^2)$의 분포를 따르는 잡음을 더하였다. C=100, $\varepsilon = 0.01$, $\sigma = 0.2$를 사용하였다. 이 그림에서 알 수 있는 사실은 전반적으로 QKASVM이 QPSVM보다 우수하다는 것을 알 수 있다. 특히 $x \in (2,3)$인 부분에서 QPSVM은 약간의 변동을 보이는 반면 QKASVM은 추정을 잘 해 주고 있다는 것을 알 수가 있다. 그러나 $x \in (0,1)$의 부분에서는 QPSVM의 결과가 좋다는 것을 알 수 있다. 몇 번의 반복된 모의 실험에서도 유사한 결과를 볼 수 있었다. 이와 같은 결과는 오차가 정규분포를 따르므로 오차제곱합 손실함수가 더 좋은 결과를 보인다는 이론을 뒷받침한다.

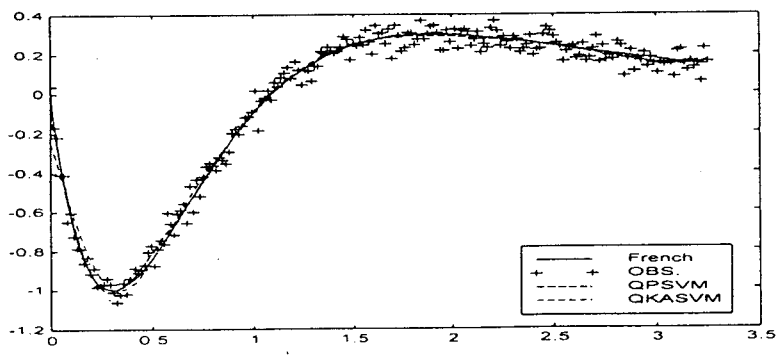

그림 5 잡음이 있는 French 곡선의 QPSVM과 QKASVM

위의 실험결과-대체적으로 QKASVM이 우수하다-를 확인하기 위해 두 알고리즘이 잡음이 더해진 곡선의 추정을 얼마나 잘 하는가를 비교하여 보았다. 모의실험은 French곡선과 싸인제곱함수, $f(x) = \sin^2(2\pi x)$에 $N(0, 0.2^2)$의 분포를 따르는 잡음을 넣은 자료를 이용하여 200회 반복(replication)을 하여 다음의 값을 얻었다.

$$\mathbf{QM} = \sum^{rep} \mathbf{Q}/rep,\ \mathbf{AM} = \sum^{rep} \mathbf{A}/rep,\ \mathbf{QSD} = sd(\mathbf{Q}),\ \mathbf{ASD} = sd(\mathbf{A})$$

여기에서
$\mathbf{Q} = \sum_{i=1}^{n} \mathbf{D}_i^2/n$, $\mathbf{A} = \sum_{i=1}^{n} |\mathbf{D}_i|/n$ 이고 sd는 표준편차인데 $\mathbf{D}_i = \mathbf{SVM}(x_i) - f(x_i)$이다. $\mathbf{SVM}(x_i)$는 QPSVM 혹은 QKASVM에 의한 추정값을 나타낸다.
표1에서 QKASVM의 값이 French 곡선을 제외한 나머지 부분 모두에서 QPSVM의 값보다 적음을 볼 수 있다. 이로 보아 QKASVM의 추정능력이 전반적으로 더 우수함을 알 수 있다. 특히 와 값이 더 적다는 것은 QKASVM의 표본변동이 더 적다는 것을 말한다. 즉, 안정된 추정값을 얻을 수 있음을 뜻한다.
이상의 모의실험의 결과에서 우리는 제안된 QKASVM의 수행능력이 함수근사에서는 조금 뒤 떨어 지지만 함수추정능력은 QPSVM보다 더 우수함을 알 수 있었다. 더욱이 계산에 소요되는 시간을 비교하였더니 같은 수준의 추정치를 계산하는데 LKASVM(LIL에 의한 KA)에서는 약 200번이상의 반복수가 필요했는데 QKASVM에서는 50번 정도의 반복수로도 충분하다는 사실을 알 수 있었다. 약 1/4의 계산시간이 소요된다는 사실이다.
다시 정리하면 제안된 QKASVM은 수행능력이 QPSVM에 비해 뒤지지 않을 뿐 아니라 빠

	French 곡선		Sinsquare 곡선	
	QKASVM	QPSVM	QKASVM	QPSVM
QM	0.001416	0.00119	0.000621	0.0006639
AM	0.0247	0.02736	0.019359	0.01985
QSD	0.000269	0.000379	0.000213	0.000345
ASD	0.003937	0.00451	0.00391	0.00526

표1**QKASVM**과**QPSVM**의 비교를 위한 통계량

르게 계산이 가능하다는 장점을 가지고 있음을 알 수 있다. 그러나 반복수가 너무 많아지거나 학습률값과 커널평활량이 제대로 설정이 되지 않을 때 가 발산하여 결과가 좋지 않다는 문제점을 지니고 있다.

4. 결론 및 향후과제

추정능력이 우수하고 계산시간을 줄일 수 있는 알고리즘은 제안하였다. SVM학습에 필요한 을 ε자동으로 조절할 수 있는 알고리즘[7]은 개발되어 있으나 C와 커널평활량을 결정하는 알고리즘은 개발이 되어있지 않다. 여기에 대한 연구가 있었으면 하는 바램이다. 그리고 제안된 알고리즘의 좀 더 나은 능력을 위해 η값을 사전에 결정할 수 있는 이론이 개발되어야 하겠다.

참 고 문 헌

1 C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition", In Data Mining and Knowledge discovery 2, Kluwer Academic Publishers, Boston, 1998.

2 C. Campbell, N. Cristianini, "Simple Learning Algorithms for Training Support Vector Machines", Dept. of Engineering Mathematics Technical Report, U. of Bristol, 1998.

[3] N. Cristianini, C. Campell and J. Shawe-Taylor, "Dynamically Adapting Kernels in Support Vector Machines", NeuroCOLT2 Technical Report, NeuroCOLT, 1998.

4 T. Evgeniou, M. Pontil and T. Poggio, "A Unified Framework and Regularizarion Networks and Support Vector Machines", MIT AI Lab., Technical Report, 1999.

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

6 E. Osunam R. Freund and F. Girosi, "Support Vector Machines : Training and Applications", MIT AI Lab., Technical Report, 1997.

7 B. Scholkopf, P. Bartlett, A. Smola and R. Williamson, "Support vector regression with automatic accuracy control", Proceedings of the 8th International Conference on Artificial Neural Networks, Perspectives in Neural Computing, Berlin, Springer Verlag, 1998.

8 X. Shao, "Model Selection Using Statistical Learning Theory", Ph. D. Thesis, U. of Minnesota, 1999.

9 A. J. Smola, B. Scholkopf, " A Tutorial in Support Vector Regression", NeuroCOLT2, Technical Report, NeuroCOLT, 1998

10 V. Vapnik, " The Nature of Statistical Learning Theory", Springer, 1995

11 V. Vapnik, "Statistical Learning Theory", Springer, 1998

12 G. Wahba and S. Wold, "A completely automatic French Curve", Communication in Statistics, 4, pp.1-17,1975.