응용통계연구 제 16권 2호, 2003년, pp. 261-271

서포트 벡터 기계에서 잡음 영향의 효과적 조절

김철응[1] 윤 민[2]

요 약

서포트 벡터 기계(Support Vector Machines, SVMs)에서의 일반화 오차의 경계는 훈련점들과 분리 초평면 사이의 최소의 거리에 의존한다. 특히, 소프트 마진 알고리즘은 목표 마진과 slack 벡터의 놈들에 의하여 경계가 결정된다. 이 논문에서는, 자료들에 있어서 잡음들에 의한 오염들을 직접적으로 고려하는 새로운 소프트 마진 알고리즘을 공식화하였다. 그리고, 수치적 예제를 통하여, 제안된 방법과 기존의 소프트 마진 알고리즘을 비교하였다.

주요용어: 서포트 벡터 기계, 소프트 마진, 일반화 오차 경계, 하드 마진, 허용오차.

1. 소개

Large 마진 분류기(margin classifier)들을 고려할 때, 가설의 복잡성은 각각의 자료들에 의하여 생기는 마진에 의하여 측정이 되는데(Cherkassky & Mulier, 1998), 만약 자료들이 잡음(noise)들이 있는 경우에는 추가적인 문제들이 발생한다. 예를 들면, 각각의 훈련 자료점들에 의해 마진을 최대화 하여 얻어진 해들은 안정적이지 못하다. 훈련집합에서 약간의 완화는 가설을 유의하게 변화시킬 수 있어 large 마진 분류기의 해들은 다소 불안정하다(Mangasarian, 2000). 이러한 문제점들은 "소프트 마진(soft-margin)"의 방법을 만들어 내게 되었는데(Shawe-Taylor & Cristianini, 2000), 이 방법의 절차는 잡음이 있는 경우에 정확도에 약간의 희생을 허용함으로써 large 마진 알고리즘의 확장을 목적으로 한다. 소프트 마진 알고리즘을 사용하는 경우에, 일반적으로 large 마진을 얻을 수 있으며, 알고리즘들은 공식화된 수리계획문제에 있어서 페널티(penalty) 모수에 매우 민감하다고 알려져 있다(Cristianini & Shawe-Taylor, 2000). 본 논문에서는, 잡음이 있는 자료에 대하여 허용 오차(allowable error)를 조절할 수 있는 새로운 방법을 소개한다. 제안한 방법은 잡음의 영향에 의한 오분류율을 감소시킬 수 있고, 또한 수치 예제들을 통하여 일반화 오차가 개선됨을 보일 것이다. 2절에서는 기존의 소프트 마진 알고리즘에 대한 일반화 오차의 경계(error bound)에 관한 내용들을 소개하고, 3절에서는 이들의 문제점과 수치 예들을 통하여 제안한 방법의 타당성을 보이고, 마지막으로 제안한 방법으로 얻을 수 있는 효과를 나타내었다.

1) (120-749) 서울시 서대문구 신촌동 134, 연세대학교 응용통계학과, 부교수
E-mail : cekim@yonsei.ac.kr
2) (120-749) 서울시 서대문구 신촌동 134, 연세대학교 응용통계학과, 시간강사
E-mail : myoon@yonsei.ac.kr

2. 소프트 마진 알고리즘에 대한 일반화 오차의 경계

우선 마진에 대한 slack 변수의 정의를 소개한다. 이 slack 변수는 각각의 자료점들이 분류에 대해 주어진 분계점(threshold)에 어떻게 만나지 못하는가를 나타낸다. 이 분계점을 본 논문에서는 목표(target) 마진이라 부른다.

정의 2.1 X를 자료공간이라 하자. X상에서 실수치 분류함수 f와 목표 마진 γ에 대하여, 만약 자료 $(\boldsymbol{x}_i, y_i) \in X \times \{-1, 1\}$가 $y_i f(\boldsymbol{x}_i) \geq \gamma$를 만족하면, $\boldsymbol{x}_i$는 정분류(correctly classified)되었다고 한다. 다른 한편으로, 만약 $y_i f(\boldsymbol{x}_i) < \gamma$ 이면, 오분류(incorrectly classified)되었다고 한다.

정의 2.2 X상에서 실수치 분류함수 f에 대하여, 각각의 함수 $f \in \mathcal{F}$에 대하여, 자료 $(\boldsymbol{x}_i, y_i) \in X \times \{-1, 1\}$의 마진 slack 변수들의 목표 마진 γ는

$$\xi_i = \xi((\boldsymbol{x}_i, y_i), f, \gamma) = \max\{0, \gamma - y_i f(\boldsymbol{x}_i)\}$$

와 같이 주어진다. 여기서 $\mathcal{F}$는 실수치 함수족(class)이다. 또한, 훈련집합 S에 대하여, ξ의 ℓ_2-놈은

$$\|\xi\|_2 = \sqrt{\sum_{(\boldsymbol{x}_i, y_i) \in S} \xi((\boldsymbol{x}_i, y_i), f, \gamma)^2}$$

와 같이 주어진다.

만약 $\xi_i > 0$이면 $(\boldsymbol{x}_i, y_i)$가 오분류임을 나타낸다. 왜냐하면 비영인 $\xi((\boldsymbol{x}_i, y_i), f, \gamma)$를 갖는 점들이 γ의 양의 마진을 얻을 수 없기 때문이다.

목표 마진 γ와 마진 slack 변수들의 ℓ_2-놈에 의하여, 선형 분류함수들에 대한 소프트 마진 알고리즘의 일반화 오차의 경계는 Shawe-Taylor 와 Cristianini(2000)의 논문에서 정리 1로 유도되었다.

정리 2.1 $\Delta > 0$이라 하자. $X \times \{-1, 1\}$상에서 고정되어 있지만 미지의 확률분포를 고려하자. 또한 X에서 원점에 대하여 반지름이 R의 공(ball)에서 서포트(support)를 가진다고 하자. 그러면 모든 $\gamma > 0$에 대하여, 크기가 ℓ인 훈련집합 S에서 $1 - \delta$의 확률로 랜덤하게 추출할 때, X상에서 $\|\mathbf{u}\| = \mathbf{1}$를 가지는 선형 분류기 $\mathbf{u}$가 0에서 분계되는 일반화 오차는

$$\varepsilon(\ell, d, \delta) = \frac{2}{\ell}\left(d \log_2\left(\frac{8e\ell}{d}\right) \log_2(32\ell) + \log_2\left(\frac{8\ell}{\delta}\right)\right),$$

와 같이 주어지고, 여기서 d는 아래와 같이 나타내어진다.

$$d = \left\lfloor \frac{64.5(R^2 + \Delta^2)\left(\frac{1 + \|\xi\|_2^2}{\Delta^2}\right)}{\gamma^2} \right\rfloor$$

또한, $\ell \geq \frac{2}{\varepsilon}$, $d \leq e\ell$이며, 오분류된 훈련점들에서의 확률은 0이다. 여기서, d는 X상에서 fat-shattering 차원이다.

정리 1은 자료들이 목표 마진 γ에 도달하지 못하는 자료들의 양에 의하여 일반화 오차의 경계가 정해지는 것을 의미한다. 오차의 경계는 slack 변수의 놈에 의하여 결정이 되는데, 일반화 능력을 향상시키기 위하여 이 놈의 값이 최소화되어야만 한다. 이 오차의 경계들은 훈련 자료들이 선형으로 분리 가능하든지, 혹은 그렇지 않은 지에는 의존하지 않고, 또한 자료들이 잡음에 오염이 된 경우에도 조절할 수 있다. 더욱이, ℓ과 δ가 고정되어 있기 때문에, 오차의 경계는 d에 대하여 단조 증가함을 아래에서 알 수 있다.

$$\begin{aligned}\left\{d\log_2\left(\frac{8e\ell}{d}\right)\right\}' &= \log_2 8e\ell - \log_2 d + \frac{1}{\log_e 2}\\ &> \log_2\left(\frac{8e\ell}{d}\right) - \left(\frac{1}{\log_2 d}\right)\\ &> 0.\end{aligned}$$

여기서 $'$은 도함수를 나타낸다.

3. 소프트 마진 알고리즘의 문제점과 새로운 방법

그림 3.1에서 보는 바와 같이 인공적인 11개의 훈련 자료점들과 이 자료들을 분류하는 네 개의 선형 분류함수 f^1, f^2, f^3, 그리고 f^4를 고려하자.

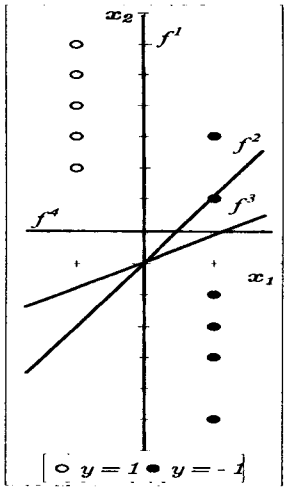

그림 3.1: 11개의 훈련 자료들과 네 개의 선형판별함수

분류함수 f^1은 자료들을 완전 분리를 하지만, 나머지 f^2, f^3, 그리고 f^4는 완전 분리를 할 수 없다. 주어진 이 네 개의 분류 함수들은 아래와 같다.

$$\begin{aligned}f^1(x_1,x_2) &= -x_1, & f^2(x_1,x_2) &= -4x_1+3x_2,\\ f^3(x_1,x_2) &= -3x_1+4x_2, & f^4(x_1,x_2) &= x_2-0.1.\end{aligned}$$

일반성을 잃지 않고, 정리1 의 수식에서 $R = 1$과 $\Delta = 1$로 두자. 그러면 여러 개의 목표 마진들에 대한 d값들은 표 1에서와 같이 나타난다. 여기서 우리는 d에 대하여 f^1, f^2, f^3, 그리고 f^4가 각각 (i) $\gamma \leq 0.25$, (ii) $0.30 \leq \gamma \leq 0.65$, (iii) $0.70 \leq \gamma \leq 1.00$, 그리고 (iv) $\gamma \geq 1.05$의 범위에서 가장 작음을 알 수 있다.

표 3.1: 주어진 목표 마진 γ에 대한 d의 값들

	d			
γ	f^1	f^2	f^3	f^4
0.05	51600	51740	52325	52684
0.10	12900	12998	13271	13416
0.15	5737	5833	6038	6135
0.20	3264	3342	3525	3597
0.25	2190	2210	2387	2445
0.30	1710	1625	1799	1851
0.35	1550	1312	1485	1532
0.40	1596	1158	1330	1375
0.45	1793	1118	1281	1326
0.50	2114	1171	1313	1361
0.55	2542	1306	1419	1467
0.60	3068	1520	1593	1638
0.65	3685	1810	1835	1875
0.70	4391	2175	2147	2182
0.75	5182	2613	2529	2559
0.80	6057	3125	2980	3005
0.85	7014	3711	3502	3522
0.90	8052	4371	4095	4108
0.95	9171	5105	4758	4766
1.00	10370	5914	5494	5495
1.05	11648	6797	6302	6296
1.10	13006	7756	7183	7170
1.15	14443	8790	8137	8117
1.20	15959	9900	9165	9137

어떤 잡음도 끼지 않은 자료를 고려하면 상대적으로 작은 γ값에서 분류가 잘 수행되었다. 다시 말하면, f^1과 같은 완전 분리 초평면들은 잡음이 없는 경우의 자료에 대해 다른 함수들보다 분류가 더욱 잘 수행된다. 그러나, 자료에 잡음이 존재하는 경우에 대하여는 상대

적으로 큰 마진 값을 고려하는 것이 좀 더 바람직하다. 그러면 γ값이 얼마나 큰 것이 적절한가에 대하여는 잡음의 영향을 어느 정도 고려할 것인가에 의존한다. 소프트 마진 알고리즘의 경우에 있어서, slack 변수 $\xi_i, i = 1, 2, \ldots, \ell$는 잡음의 영향을 고려하기 위하여 소개되었다.

SVM에서 기존의 소프트 마진 문제를 ℓ_1-놈을 가지고 공식화하면 아래와 같다(Shawe-Taylor & Cristianini, 2000).

$$\begin{aligned} \underset{\boldsymbol{w},w_0,\xi_i}{\text{minimize}} \quad & \|\boldsymbol{w}\|_1 + C\sum_{i=1}^{\ell}\xi_i \\ \text{subject to} \quad & y_i\left(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + w_0\right) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \ \ i = 1, \ldots, \ell, \end{aligned} \tag{S}$$

여기서 C는 slack 변수에 대한 가중치 모수이고 $\|\boldsymbol{w}\|_1$는 $\sum_{i=1}^{\ell}|w_i|$를 나타낸다.

최적 분리 초평면은 주어진 모수 C에 대하여 위의 문제 (S)를 풀면 구하여진다. 일반적으로 벌칙상수 C값의 결정은 주어진 $\|\omega\|_2$의 값을 가지는 $\|\boldsymbol{\xi}\|_1$을 최소화하여 얻어질 수 있다. 실제로, 모수 C는 잡음의 영향에 따라서 경험적으로 선택된다. 예를 들면, 만약 잡음의 영향이 크다면 적절히 작은 C를 택한다.

소프트 마진 알고리즘(S)를 이용하여, 앞 절에 사용된 자료들에 대한 분리 평면들을 구하여 보았다. 이 경우에 있어서, $C = 0.0001, 0.001, 0.01, 0.1, 1.0, 2.0$에 대하여 그림 3.2는 여러 가지의 C값에 대한 최적 분리 초평면을 나타낸다. 비록 모수 C가 변하지만, 우리는 단지 두 종류의 분리 초평면만을 얻을 수 있다. 즉, 기존의 소프트 마진 알고리즘에 의하여 벌칙상수 C가 어떤 값을 가지든지 f^2와 f^3와 같은 비수직(non-vertical)과 비수평(non-horizontal)인 분리 평면은 얻을 수가 없다. 이 사실은 기존의 소프트 마진 알고리즘의 방법으로 잡음의 영향을 직접적으로 고려할 수 없음을 나타내고 있다.

본 논문에서는 위에서 언급한 문제를 극복하고, 잡음의 영향을 직접적으로 평가 할 수 있는 새로운 방법을 제안하고자 한다.

$$\begin{aligned} \underset{\boldsymbol{w},w_0,\xi_i}{\text{minimize}} \quad & \|\boldsymbol{w}\|_1 + C\sum_{i=1}^{\ell}\xi_i \\ \text{subject to} \quad & y_i\left(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + w_0\right) \geq 1 - \xi_i, \\ & 0 \leq \xi_i \leq \xi_{\max}, \ \ i = 1, \ldots, \ell, \end{aligned} \tag{$S_{\xi_{\max}}$}$$

여기서 C는 slack 변수에 대한 가중치 모수이고 $\xi_{\max}$는 주어진 고정된 상수이다. 위의 공식 에서, 마진 slack 변수의 상한값 $\xi_{\max}$를 설정하였는데, 이는 허용오차를 의미한다. 우리는 이를 이용하여 잡음의 영향을 직접 조절할 수 있다. 예를 들면, 상대적으로 큰 $\xi_{\max}$의 값에 대하여, 가중치 모수 C만을 고려하는 방법보다 $\xi_{\max}$를 동시에 이용하는 것이 훨씬 더 직접적으로 잡음의 영향을 고려한 분리 초평면을 얻을 수 있다.

그림 3.3은 제안된 방법에 의하여 얻어진 최적 분리 초평면을 나타낸다. $C = 1.0$의 경우를 예를 들어 제안한 새로운 방법 $(S_{\xi_{\max}})$와 기존의 (S)를 비교하자. 제안한 방법에서 $\xi_{\max}$값

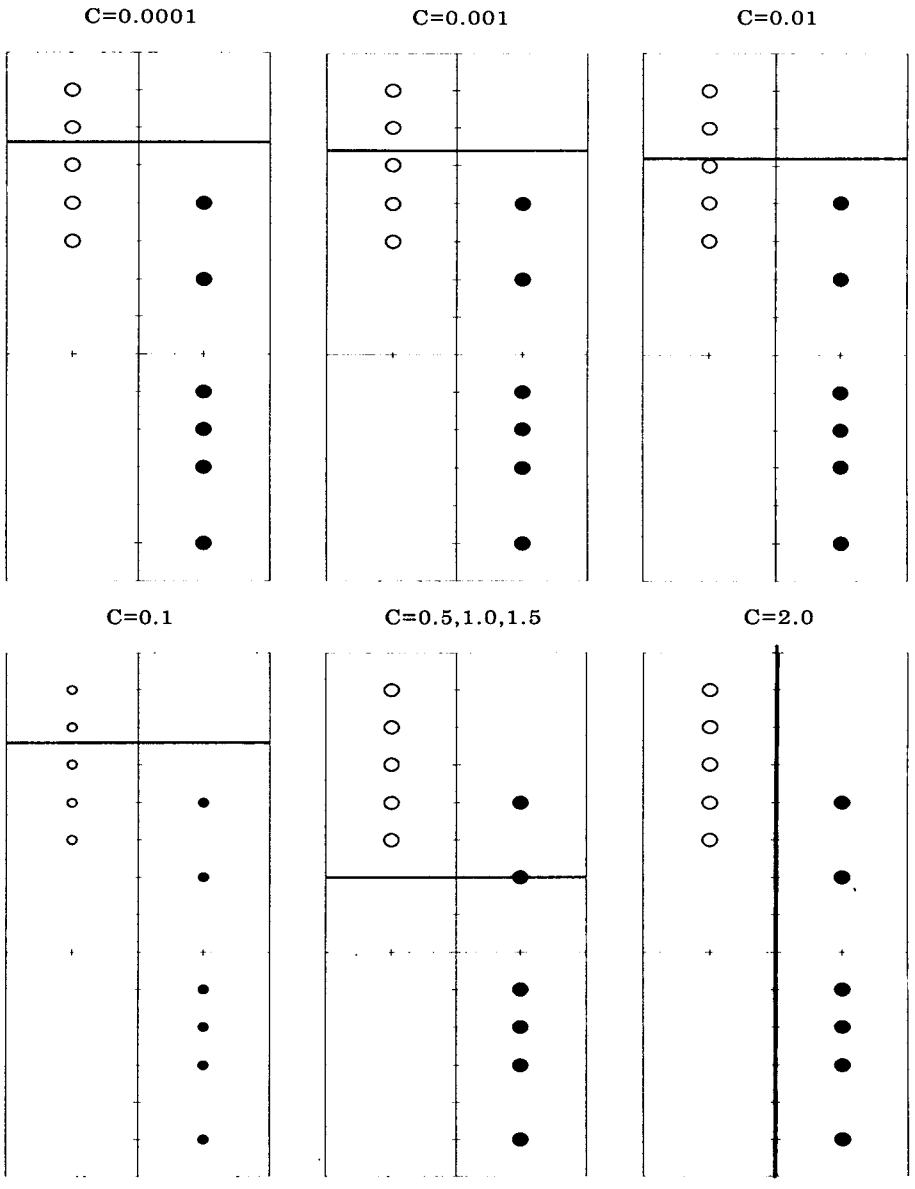

그림 3.2: (S)에서 여러 가지 C값들의 변화에 의한 분리 초평면들

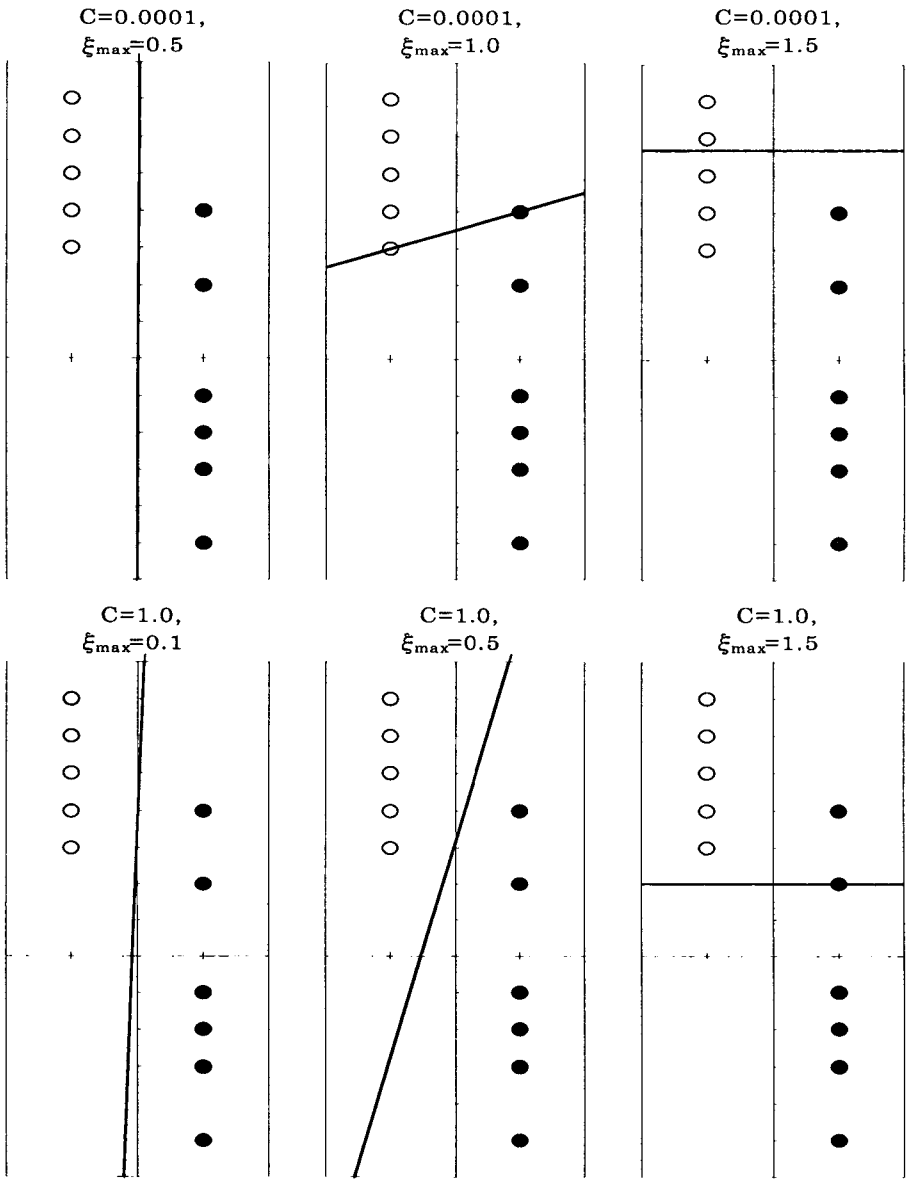

그림 3.3: $(S_{\xi_{max}})$에서 여러 가지 C값들의 변화에 의한 분리 초평면들

들의 변화에 의하여 여러 형태의 분리 초평면들을 얻을 수 있으나, 반면 (S)문제를 풀면 그림 3.2에서와 같이 단지 수평의 최적 분리 초평면 밖에 나타나지 않는다. $\xi_{\max}$값의 조절에 의하여, 얼마나 많은 잡음의 영향이 고려될지의 정도에 따라서 적절한 분리 초평면들을 얻을 수 있다. 왜냐하면 $\xi_{\max}$자체가 허용오차의 수준으로서 역할을 하고, 실제 문제에 적용하는 경우에는 잡음의 영향이 작다면 가능한 작은 $\xi_{\max}$값을 택하고, 잡음의 영향이 크다면 큰 값을 선택함으로서 우리는 그 값을 경험적으로 쉽게 결정할 수 있기 때문이다.

제안된 방법의 효율성을 알아보기 위하여, 잘 알려진 Fisher의 붓꽃 자료에 난수를 발생시켜서 인위적으로 잡음을 첨가하여 사용하였다. 이 경우, 우리는 이진분류의 문제를 고려하기 때문에, 사용한 자료에서 직관적으로 완전 분리가 되는 Setosa종은 제거를 한 후, Versicolor종과 Verginica의 두 종만을 사용하였다. 여러 가지의 C값들에 대하여 실험을 실행하여 대부분의 경우에서 기존의 방법보다 우수한 결과를 얻을 수 있었다. 그 결과는 아래의 표 3.2에서 (S)와 $(S_{\xi_{\max}})$에 대한 정분류율을 나타내었다. 잡음의 값이 각각 0.5, 1.5, 그리고 2.0은 각각 균일분포 $U[-0.5, 0.5]$, $U[-1.5, 1.5]$, 그리고 $U[-2.0, 2.0]$를 따르는 난수이다. 이 표의 값들은 100번의 실험을 한 평균값이다.

표 3.2: 붓꽃 자료를 이용한 (S)와 $(S_{\xi_{\max}})$에 대한 정분류율

노이즈	C값	(S)	$\xi_{\max}$		
			1.50	1.75	2.00
	0.10	82.33	83.67	84.00	82.67
0.50	1.00	83.67	85.00	**86.00**	85.00
	10.0	83.67	84.33	84.67	85.33

노이즈	C값	(S)	$\xi_{\max}$		
			1.50	2.00	2.50
	0.10	82.00	**85.33**	83.67	82.00
1.50	1.00	83.33	85.00	84.33	85.00
	10.0	82.33	84.33	83.33	84.00

노이즈	C값	(S)	$\xi_{\max}$		
			2.00	2.50	3.00
	0.10	72.33	72.33	73.00	72.33
2.00	1.00	74.00	74.00	**75.33**	74.33
	10.0	73.33	73.67	75.00	74.67

또 다른 수치 예로서 알래스카산과 캐나다산 연어들이 담수와 해수에서 성장 영역의 지름을 측정한 자료를 사용하였다 (Johnson & Wichern, 1992). 이 자료의 경우에도 마찬가지로 인위적으로 난수를 생성하여 원래의 자료에 잡음으로 추가하였고, 벌칙상수가 $C = 0.1$, $C = 1.0$ 그리고 $C = 10.0$인 경우에 대하여 각각 실험을 하였다. 그 결과, 대부분의 경우에

서 기존의 방법에 의한 결과보다 우수한 정분류율을 나타내었다. 아래의 표 3.3의 경우에서 이들의 값을 표로 나타내어 비교하였다. 마찬가지로 이 표에서 주어진 값들은 100번의 실험을 시행한 평균값이다. 잡음이 0인 경우는 원 자료이고, 잡음의 값이 각각 0.5, 1.5, 그리고 2.0은 각각 균일분포 $U[-0.5, 0.5]$, $U[-1.5, 1.5]$, 그리고 $U[-2.0, 2.0]$를 따르는 난수이다. 일반적으로, 적절한 벌칙상수 C와 $\xi_{\max}$값을 결정하기는 어렵다.특히 $\xi_{\max}$값의 경우에 잡음의 영향이 크다면 큰 값의 $\xi_{\max}$를 사용하고, 그 반대의 경우라면 작은 값을 사용하는데 이는 분석자가 경험적으로 선택하게 된다. 그러나, 아래의 표 3.2와 표 3.3에서 나타내는바와 같이 기존의 소프트 마진 알고리즘보다는 제안한 $(S_{\xi_{\max}})$에 의하여 더욱 개선된 결과를 얻을 수 있었다.

표 3.3: 연어 자료를 이용한 (S)와 $(S_{\xi_{\max}})$에 대한 정분류율

노이즈	C값	(S)	$\xi_{\max}$		
			1.25	1.50	1.75
	0.10	94.33	94.40	94.90	**95.33**
0.00	1.00	93.88	94.35	94.78	94.88
	10.0	93.78	94.28	94.75	94.33
노이즈	C값	(S)	$\xi_{\max}$		
			1.50	2.00	2.50
	0.10	87.67	**91.00**	88.67	87.67
0.50	1.00	89.33	**91.00**	89.67	89.33
	10.0	89.67	**91.00**	90.00	90.33
노이즈	C값	(S)	$\xi_{\max}$		
			1.50	2.00	2.50
	0.10	87.00	88.67	87.67	87.00
1.50	1.00	86.67	88.67	89.00	88.33
	10.0	86.67	88.67	**89.33**	89.00
노이즈	C값	(S)	$\xi_{\max}$		
			2.50	2.75	3.00
	0.10	76.67	76.67	76.67	76.67
2.00	1.00	77.67	78.67	78.00	78.00
	10.0	78.33	**79.56**	78.67	78.67

4. 결론

본 논문에서, 잡음이 있는 자료에 대하여 허용오차를 조절하기 위하여 소프트 마진 알

고리즘의 새로운 공식을 제안하였다. 제안된 방법은 잡음의 영향을 직접 조절할 수 있는의 ξ_{max} 값들을 조절함으로서 분리성능이 향상된 분리 초평면들을 쉽게 얻을 수 있음을 알 수 있었다.

참고문헌

[1] Bartlett, P. and Shawe-Taylor, J. (1999). Generalization Performance of Support Vector Machines an Other Pattern Classifiers. *Advances in Kernel Methods-Support Vector learning*,(edited by Schölkopf, B., Burges, C. J. C., and Smola, A.), 43-54. MIT Press.

[2] Cherkassky, V. and Mulier, F. (1998). *Learning from Data : Concepts, Theory, and Methods*, John Wiley & Sons.

[3] Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and other kernel-based learning methods*, Cambridge University Press.

[4] Johnson, R.A. and Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis*(3rd ed.), Prentice-Hall, Inc.

[5] Mangasarian, O.L. (2000). Generalized Support Vector Machines. *Advances in Large Margin Classifiers*, (edited by Smola, A., Bartlett, B., Schölkopf, B., and Schuurmans, D.), 135-146. MIT Press.

[6] Shawe-Taylor, J. and Cristianini, N. (2000). On the generalization of Soft margin Algorithms, *NeuroCOLT2 Technical Report Series*, NC-TR-2000-082.

[7] Smola, A.J. and Schölkopf, B. (1998). A Tutorial on Support Vector Regression, *NeuroCOLT2 Technical Report*, NeuroCOLT.

[8] Vapnik, V. (1998). *Statistical Learning Theory*, John Wiley & Sons.

[2002년 10월 접수, 2003년 3월 채택]

Support Vector Machines Controlling Noise Influence Effectively

Chul Eung Kim [1] Min Yoon [2]

ABSTRACT

Support Vector Machines (SVMs) provide a powerful performance of the learning system. Generally, SVMs tend to make overfitting. For the purpose of overcoming this difficulty, the definition of soft margin has been introduced. In this case, it causes another difficulty to decide the weight for slack variables reflecting soft margin classifiers. Especially, the error of soft margin algorithm can be bounded by a target margin and some norms of the slack vector.

In this paper, we formulate a new soft margin algorithm considering the bound of corruption by noise in data directly. Additionally, through a numerical example, we compare the proposed method with a conventional soft margin algorithm.

Keywords: Support vector machines(SVMs); soft margin, generalization error bound; hard margin; allowable error

1) Associate Professor, Dept. of Applied Statistics, Yonsei University.
E-mail : cekim@yonsei.ac.kr
2) Par-time lecturer, Dept. of Applied Statistics, Yonsei University.
E-mail : myoon@yonsei.ac.kr