응용통계연구 제 17권 1호, 2004년, pp. 75-88

서포트 벡터 기계에서 TOTAL MARGIN을 이용한 일반화 오차 경계의 개선

윤 민[1)]

요 약

서포트 벡터 기계(Support Vector Machines, SVMs) 알고리즘은 표본 점들과 분리 초평면 사이의 최소 거리를 최대화하는 것에 관심을 가져왔다. 본 논문은 모든 데이터 점들과 분리 초평면 사이의 거리들을 고려하는 total margin을 제안한다. 본 논문에서 제안하는 방법은 기존의 서포트 벡터 기계 알고리즘을 확장하고, 일반화 오차 경계를 개선하게 된다. 새롭게 제안하는 total margin 알고리즘이 기존 방법들과의 비교를 통하여 더욱 우수한 수행능력을 가지고 있음을 수치 예제들을 통하여 확인할 수 있다.

주요용어: 서포트 벡터 기계, 일반화 오차 경계, Soft Margin, Surplus 변수, Total Margin.

1. 소개

서포트 벡터 기계들은 많은 유용한 특징들과 실제적인 수행능력의 장점들을 가지고 있기 때문에 기계학습에 있어서 더욱 더 많은 인기를 얻고 있다. 서포트 벡터 기계의 주요한 아이디어들 중의 하나는 표본점들에서부터 분리 초평면 사이의 최소 거리를 최대화함으로서 얻어지는 최대화(maximal) margin을 갖는 선형분류기(linear classifier)에 근거하고 있다.

먼저, 훈련 데이터 집합 $S = ((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_\ell, y_\ell)) \in (X \times \{-1, 1\})^\ell$를 고려하자. 여기서 X는 입력 공간이다. 각각의 데이터들은 선형함수 $f(\boldsymbol{x}) = \langle \boldsymbol{w} \cdot \boldsymbol{x} \rangle + w_0$의 부호로 분류되어진다고 가정하자. 기존의 최대화 margin(hard margin) 알고리즘은 아래와 같이 공식화할 수 있다.

$$\begin{aligned} \underset{\boldsymbol{w}, w_0}{\text{minimize}} \quad & \langle \boldsymbol{w} \cdot \boldsymbol{w} \rangle \\ \text{subject to} \quad & y_i \left(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + w_0 \right) \geq 1, \ \ i = 1, \ldots, \ell. \end{aligned}$$

이 알고리즘을 통하여, 최적의 분리 초평면은 표본점들로부터 최소의 거리를 최대화함으로서 얻을 수 있다. 만약 데이터들이 잡음에 의하여 오염된 자료들이라면 실제로 완전 분리 초평면이 존재하지 않을 수도 있다. 이러한 어려움을 개선하기 위하여, 데이터 점들에

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

대한 margin의 측도를 완화하기 위하여 slack 변수들이 도입되었다. 이 soft margin 알고리즘은 아래와 같이 공식화할 수 있다:

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

여기서 C는 slack 변수에 대한 가중치 모수이고 ξ_i^-는 slack 변수이다.

본 논문에서 margin의 측도로써 모든 데이터 점들과 분리 초평면 사이의 거리를 고려하는 surplus 변수들을 정의한다. 오분류(misclassified)된 데이터(즉, slack 변수)와 정분류(correctly classified)된 데이터(즉, surplus 변수)를 동시에 고려하는 total margin 알고리즘을 제안한다. 여기서 total margin은 모든 데이터 점들과 분리 초평면 사이의 거리의 합으로서 표현된다. 그리고 slack 변수들을 최소화하고 surplus 변수들을 최대화함으로서 total margin에 대한 일반화 오차 경계를 감소할 수 있다는 사실을 증명할 것이다. 마지막으로, 수치 예제들을 통하여 기존의 soft margin 알고리즘과 total margin 알고리즘을 비교한다.

2. Total Margin을 이용한 일반화 오차의 경계

이 절에서, 먼저 본 논문에서 나타나는 중요한 여러 개의 기존의 결과들을 대략적으로 생각하자. 먼저 분류기들(classifiers)의 측도로써 제공되는 "margin"은 표본 점들과 분리 함수 사이의 거리를 의미한다.

정의 2.1 분계점(thresholding) 0에 의한 분류에 대하여, 입력공간 X상에서 실수치 함수족(class) $\mathcal{F}$를 고려하자. 함수 $f\in\mathcal{F}$에 의한 자료 $(\boldsymbol{x}_i,y_i)\in X\times\{-1,1\}$의 margin은

$$\gamma_i = y_i f(\boldsymbol{x}_i)$$

와 같이 정의되는 양으로 나타난다.

표준 서포트 벡터 기계(hard margin) 알고리즘은 margin γ_i $(i=1,2,\ldots,\ell)$들의 최소치를 최대화한다. 다른 한편으로, 잡음에 대하여 강건한(robust) 분류를 하기 위하여 slack 변수들을 도입함으로서 soft margin 알고리즘이 개발되었다(Cortes, 1995). 아래의 정의 2.2와 정리 2.1은 soft margin 알고리즘의 주요한 개념들이다.

정의 2.2 (Slack 변수): 분계점(thresholding) 0에 의한 분류에 대하여, 입력공간 X상에서 실수치 함수족(class) $\mathcal{F}$를 고려하자. 함수 $f\in\mathcal{F}$와 목표(target) margin γ에 대하여, 데이터 $(\boldsymbol{x}_i,y_i)\in X\times\{-1,1\}$의 *margin slack* 변수는

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

의 양으로 주어진다. 함수 f와 목표 margin γ에 의하여 훈련집합 S의 *margin slack* 벡터는 margin slack 변수들의 벡터들로

$$\boldsymbol{\xi}^- = \boldsymbol{\xi}^-(S, f, \gamma) = (\xi_1^-, \xi_2^-, \ldots, \xi_\ell^-)$$

와 같이 나타내어진다.

그림 2.1에서 보는바와 같이, ξ_i^-는 점 $(\boldsymbol{x}_i, y_i)$에서 목표 margin γ에 도달하지 못한 양으로 나타낸다. 아래의 정리 2.1은 soft margin 알고리즘의 일반화 오차 경계를 보여준다.

정리 2.1 (Shawe-Taylor 와 Cristianini, 2000): $\Delta > 0$이라 하자. $X \times \{-1, 1\}$상에서 고정되어 있는 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 $\|\boldsymbol{u}\| = 1$를 가지는 선형 분류기(classifier) $\boldsymbol{u}$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

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

이고 $\ell \geq \frac{2}{\varepsilon}$, $d \leq e\ell$이며, 오분류된 훈련 점들에서의 확률은 0이다.

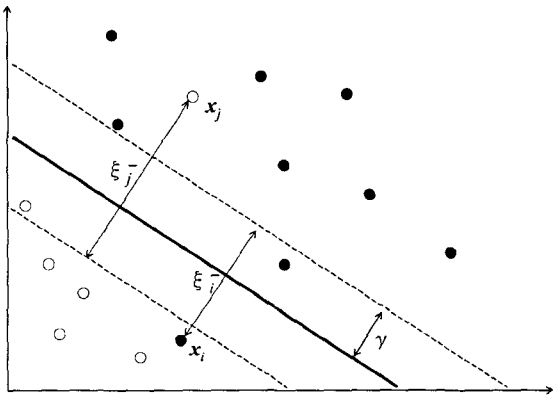

그림 2.1: slack 변수들

잡음이 있는 데이터의 경우, soft margin은 slack 변수들을 도입하여 정확성은 조금 감소되지만 hard margin 알고리즘을 개선하는데 목적이 있다. 여기서 우리는 새롭게 "surplus

변수들"을 소개한다. 개념적으로, surplus 변수들은 slack 변수들의 반대의 효과를 가진다. Slack 변수들이 오분류된 데이터 점들의 측정량인데 비하여, surplus 변수들은 정분류된 데이터 점들의 측정량이다.

정의 2.3 (Surplus 변수): X상에서 실수치 분류함수 f에 대하여, 함수 $f \in \mathcal{F}$와 목표 margin γ에 대하여, 자료 $(\boldsymbol{x}_i, y_i) \in X \times \{-1, 1\}$의 *margin surplus* 변수는

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

로 정의한다. 여기서 $\mathcal{F}$는 실수치 함수족이다.

함수 f와 목표 margin γ에 의하여 훈련집합 S의 *margin surplus* 벡터는

$$\boldsymbol{\xi}^+ = \boldsymbol{\xi}^+ (S, f, \gamma) = (\xi_1^+, \xi_2^+, \ldots, \xi_\ell^+)$$

와 같이 정의한다. 특히

$$\frac{1}{\boldsymbol{\xi}^+ + \mathbf{1}} = \left(\frac{1}{{\xi_1}^+ + 1}, \frac{1}{{\xi_2}^+ + 1}, \ldots, \frac{1}{{\xi_\ell}^+ + 1}\right)$$

라 두면, $\boldsymbol{\xi}^+$의 ℓ_2-놈(norm)은

$$\left\| \frac{1}{\boldsymbol{\xi}^+ + \mathbf{1}} \right\|_2 = \sqrt{\sum_{(\boldsymbol{x}_i, y_i) \in S} \left(\frac{1}{\xi^+((\boldsymbol{x}_i, y_i), f, \gamma) + 1} \right)^2}$$

와 같이 정의할 수 있다.
그림 2.2에서 ξ_i^+는 목표 margin γ에서 정분류된 점 $(\boldsymbol{x}_i, y_i)$들 사이의 거리를 나타낸다.

Slack 변수들과 surplus 변수들을 도입한 오차의 경계를 얻기 위하여, 우리는 먼저 Shawe-Taylor 와 Cristianini(2000)논문에서 소개된 정의를 아래에서 간략하게 소개한다.

정의 2.4 $L(X)$는 셀 수 있는 서포트(support) $\mathrm{supp}(f)$를 가지는 X상에서 실수치 함수들 f의 집합이라 하자. 즉, $L(X)$에서 함수들은 단지 가산적으로(countably) 많은 점들에 대하여 0이 아니다. 두 종류의 놈을 고려하자. 우선 ℓ_2-놈 $\|f\|_2$는

$$\|f\|_2^2 = \sum_{\boldsymbol{x} \in \mathrm{supp}(f)} f(\boldsymbol{x})^2 < \infty$$

와 같이 정의되고, 반면에 ℓ_1-놈은

$$\|f\|_1 = \sum_{\boldsymbol{x} \in \mathrm{supp}(f)} |f(\boldsymbol{x})| < \infty$$

와 같이 주어진다. 두 함수들 f, g의 내적은

$$\langle f \cdot g \rangle = \sum_{\boldsymbol{x} \in \mathrm{supp}(f)} f(\boldsymbol{x}) g(\boldsymbol{x})$$

에 의하여 정의된다. 이 공간은 덧셈과 스칼라 곱에 대하여 닫혀있음을 쉽게 확인할 수 있다.

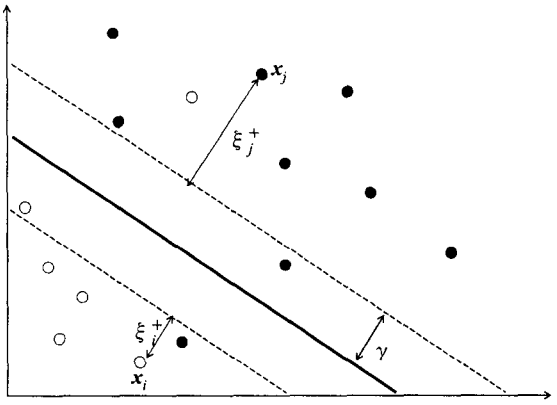

그림 2.2: surplus 변수들

정의 2.5 고정된 $\Delta > 0$에 대하여 입력공간 X를 내적공간 $X \times L(X)$로의 embedding의 정의는 아래와 같다.

$$\tau_\Delta : \boldsymbol{x} \mapsto X_\Delta = (\boldsymbol{x}, \Delta\delta_{\boldsymbol{x}}),$$

여기서 $\delta_{\boldsymbol{x}} \in L(X)$는

$$\delta_{\boldsymbol{x}}(z) = \begin{cases} 1, \text{if } \boldsymbol{x} = \boldsymbol{z} \\ 0, \text{if otherwise} \end{cases}$$

와 같이 정의된다.

Shawe-Taylor 와 Cristianini(2000)의 방법과 유사한 방법으로 주어진 실수치 함수 f와 훈련 데이터 집합 S에 대하여 보조함수 $g_f \in L(X)$를 아래와 같이 구성한다:

$$g_f = \frac{1}{\Delta} \sum_{(\boldsymbol{x},y)\in S} y \frac{\xi^-((\boldsymbol{x},y), f, \gamma)}{\xi^+((\boldsymbol{x},y), f, \gamma) + 1} \delta_{\boldsymbol{x}}$$

이 보조 함수는 (f, g_f)가 아래와 같이 훈련 데이터 집합 S상에서 목표 margin γ를 얻을 수 있음을 보장할 수 있다. 위의 보조함수 g_f에 대하여, 아래의 보조정리를 얻을 수 있다.

보조정리 2.1 임의의 훈련집합 S, 실수치 함수 f와 margin γ에 대하여, 함수 (f, g_f)는 아래의 두 가지 성질들을 만족한다.

(1) $m((f, g_f), \tau_\Delta(S)) \geq \gamma$이고,

(2) $(\boldsymbol{x}, y) \neq S$에 대하여, $(f, g_f)(\tau_\Delta(\boldsymbol{x}), y) = f(\boldsymbol{x})$이다.

증명: 1. 훈련점 $(\tau_\Delta(\boldsymbol{x}), y) \in \tau_\Delta(S)$를 함수 (f, g_f)의 margin에 적용하면,

$$\begin{aligned} y(f, g_f)\tau_\Delta(\boldsymbol{x}) &= yf(\boldsymbol{x}) + \frac{y}{\Delta} \sum_{(\boldsymbol{x}', y') \in S} y' \frac{\xi^-((\boldsymbol{x}', y'), f, \gamma)}{\xi^+((\boldsymbol{x}', y'), f, \gamma) + 1} \langle \delta'_{\boldsymbol{x}}, \Delta\delta_{\boldsymbol{x}} \rangle \\ &= yf(\boldsymbol{x}) + \frac{\xi^-((\boldsymbol{x}', y'), f, \gamma)}{\xi^+((\boldsymbol{x}', y'), f, \gamma) + 1} \\ &\geq \gamma \end{aligned}$$

를 얻는다. 정의 2.1, 2.2, 그리고 2.3으로부터 만약 $(\boldsymbol{x}, y)$이 정분류 되었다면,

$$\begin{aligned} yf(\boldsymbol{x}) &\geq \gamma \\ \xi^-((\boldsymbol{x}, y), f, \gamma) &= 0 \\ \xi^+((\boldsymbol{x}, y), f, \gamma) &= yf(\boldsymbol{x}) - \gamma \geq 0 \end{aligned}$$

이고, 그 외의 경우는 즉, $(\boldsymbol{x}, y)$이 오분류 되었다면,

$$\begin{aligned} yf(\boldsymbol{x}) &< \gamma \\ \xi^+((\boldsymbol{x}, y), f, \gamma) &= 0 \\ \xi^-((\boldsymbol{x}, y), f, \gamma) &= \gamma - yf(\boldsymbol{x}) > 0 \end{aligned}$$

이다. 따라서 첫 번째 성질은 성립한다.

2. 함수 (f, g_f)를 훈련 집합에 속하지 않는 점 $(\tau_\Delta(\boldsymbol{x}), y) \notin \tau_\Delta(S)$에 적용하면, $(\boldsymbol{x}', y') \in S$와 $\langle \delta'_{\boldsymbol{x}} \cdot \delta_{\boldsymbol{x}} \rangle = 0$에 대하여 아래의 사실을 알 수 있다.

$$\begin{aligned} \langle g_f \cdot \delta_{\boldsymbol{x}} \rangle &= \sum_{(\boldsymbol{x}', y') \in S} y' \frac{\xi^-((\boldsymbol{x}', y'), f, \gamma)}{\xi^+((\boldsymbol{x}', y'), f, \gamma) + 1} \langle \delta_{\boldsymbol{x}'}, \Delta\delta_{\boldsymbol{x}} \rangle \\ &= 0. \end{aligned}$$

따라서, 이 보조 정리의 두 번째 성질은 훈련 집합에 속하지 않는 점 $\boldsymbol{x}$에 대하여 (f, g_f)의 작용(action)은 원래의 함수와 정확하게 일치한다. 즉, $(f, g_f)\tau_\Delta(\boldsymbol{x}) = f(\boldsymbol{x})$이다. □

마지막으로 surplus 변수들과 slack 변수들을 가지는 일반화 오차 경계는 아래의 정리에서 나타낸다. 이것은 정리 2.1을 확장한 형태로 생각할 수 있다.

정리 2.2 $\Delta > 0$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 $\|\boldsymbol{u}\| = 1$를 가지는 선형 분류기(classifier) $\boldsymbol{u}$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

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

이고, $\ell \geq \frac{2}{\epsilon}$, $d \leq e\ell$이며, 오분류된 훈련 점들에서의 확률은 0이다.

증명: 내적공간 $X \times L(X)$상에서 고정된 함수 τ_Δ와 확장된 선형 함수

$$\boldsymbol{u}' = (\boldsymbol{u}, g_{\boldsymbol{u}})$$

를 고려하자.

보조정리 1의 첫 번째 성질에 의하여, $\boldsymbol{u}'$는 $\tau_\Delta(S)$상에서 목표 margin γ를 가진다. 즉,

$$m(\boldsymbol{u}', \tau_\Delta(S)) \geq \gamma$$

를 만족하고, 반면에 새로운 자료들에서 이것의 실행(action)은 $\boldsymbol{u}$와 부합된다. 게다가, $\boldsymbol{u}'$는 공간 $X \times L(X)$상에서 선형 함수이기 때문에,

$$\hat{\boldsymbol{u}} = \frac{\boldsymbol{u}'}{\|\boldsymbol{u}'\|}$$

와 같은 함수를 만들 수 있다. 그러면, $\hat{\boldsymbol{u}}$의 놈은 1이고, 아래의

$$\begin{aligned} m(\hat{\boldsymbol{u}}, \tau_\Delta(S)) &= m\left(\frac{\boldsymbol{u}'}{\|\boldsymbol{u}'\|}\ ,\ \tau_\Delta(S)\right) \\ &\geq \frac{\gamma}{\|\boldsymbol{u}'\|} = \frac{\gamma}{\|(\boldsymbol{u}, g_{\boldsymbol{u}})\|} \\ &= \frac{\gamma}{\sqrt{\|\boldsymbol{u}'\|^2 + \sum_{(\boldsymbol{x}_i, y_i)\in S}\left(\frac{\xi^-((\boldsymbol{x}_i,y_i),f,\gamma)}{\xi^+((\boldsymbol{x}_i,y_i),f,\gamma)+1}\right)^2 \frac{1}{\Delta^2}}} \\ &\geq \frac{\gamma}{\sqrt{1+\|\boldsymbol{\xi}^-\|_2^2 \left\| \frac{1}{\boldsymbol{\xi}^+ +1} \right\|_2^2 \frac{1}{\Delta^2}}} \end{aligned}$$

을 만족한다. 이것은 또한 $(\boldsymbol{x}, y) \neq S$에 대한 $\boldsymbol{u}$의 판별과 유사하다. □

$\tau_\Delta : \boldsymbol{x} \mapsto (\boldsymbol{x}, \Delta\delta_{\boldsymbol{x}})$이기 때문에, $\tau_\Delta(\boldsymbol{x})$의 분포에 대한 서포트는 반지름이 $\sqrt{R^2+\Delta^2}$인 공의 내부에 포함되는 사실에 주목해야 한다.

정리 2.2에서 알 수 있는바와 같이, $\varepsilon(\ell, d, \delta)$는 ℓ과 δ가 고정되어 있기 때문에 d에 대하여 단조적으로 증가한다(김철응, 윤민, 2003). 게다가, d는 목표 margin과 slack 변수들과 surplus 변수들의 크기에 의존한다. 끝으로, 일반화 오차는 slack 변수들과 surplus 변수들의 놈에 의하여 주어지고, 이것은 일반화 오차를 줄이기 위하여 $\|\boldsymbol{\xi}^-\|_2$와 $\left\| \frac{1}{\boldsymbol{\xi}^+ +1} \right\|$의 크기가 최소화가 되어야 함을 의미한다.

3. Total Margin 알고리즘의 공식화

Slack 벡터를 최소화하고 surplus 벡터를 최대화하기 위하여, 아래의 최적화 문제 (T)를 공식화하면:

$$\begin{aligned} \underset{\boldsymbol{w},w_0,\boldsymbol{\xi}^-,\boldsymbol{\xi}^+}{\text{minimize}} \quad & \langle \boldsymbol{w}\cdot\boldsymbol{w}\rangle + C_1\sum_{i=1}^{\ell}\xi_i^- - C_2\sum_{i=1}^{\ell}\xi_i^+ \\ \text{subject to} \quad & y_i\left(\langle \boldsymbol{w}\cdot\boldsymbol{x}_i\rangle + w_0\right) \geq 1-\xi_i^- + \xi_i^+, \\ & \xi_i^- \geq 0,\ \xi_i^+ \geq 0,\ i=1,\ldots,\ell, \end{aligned} \tag{T}$$

와 같고, 여기서 C_1와 C_2는 $C_1 > C_2$를 만족시키는 값으로 선택된다. $C_1 > C_2$의 조건은 ξ_i^-와 ξ_i^+중에서 적어도 하나의 값이 영이 된다는 사실을 보장한다는 것을 알 수 있다(Freed와 Glover, 1981). 우선, total margin 알고리즘을 만드는 일차문제 (T)의 쌍대(dual) 문제 (T_D)를 알아본다. 문제 (T)에 대한 라그랑쥬(Lagrangian) 함수는

$$\begin{aligned} L(\boldsymbol{w},w_0,\boldsymbol{\xi}^-,\boldsymbol{\xi}^+,\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}) = & \ \frac{1}{2}\langle \boldsymbol{w}\cdot\boldsymbol{w}\rangle + C_1\sum_{i=1}^{\ell}\xi_i^- - C_2\sum_{i=1}^{\ell}\xi_i^+ \\ & - \sum_{i=1}^{\ell}\alpha_i\left[y_i(\langle \boldsymbol{w}\cdot\boldsymbol{x}_i\rangle + w_0) - 1 + \xi_i^- - \xi_i^+\right] - \sum_{i=1}^{\ell}\beta_i\xi_i^- - \sum_{i=1}^{\ell}\gamma\xi_i^+ \end{aligned}$$

와 같고, 여기서 $\alpha_i \geq 0$, $\beta_i \geq 0$, 그리고 $\gamma_i \geq 0$이다. 문제 (T)의 쌍대 문제를 얻기 위하여, $\boldsymbol{w}$, w_0, $\boldsymbol{\xi}^-$, 그리고 $\boldsymbol{\xi}^+$ 각각에 대하여 미분을 하면 아래의 조건들을 얻는다:

$$\begin{aligned} \frac{\partial L(\boldsymbol{w},w_0,\boldsymbol{\xi}^-,\boldsymbol{\xi}^+,\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \boldsymbol{w}} &= \boldsymbol{w} - \sum_{i=1}^{\ell}\alpha_i y_i \boldsymbol{x}_i = \boldsymbol{0}, \\ \frac{\partial L(\boldsymbol{w},w_0,\boldsymbol{\xi}^-,\boldsymbol{\xi}^+,\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \boldsymbol{\xi}_i^-} &= C_1 - \alpha_i - \beta_i = 0, \\ \frac{\partial L(\boldsymbol{w},w_0,\boldsymbol{\xi}^-,\boldsymbol{\xi}^+,\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \boldsymbol{\xi}_i^+} &= -C_2 + \alpha_i - \gamma_i = 0, \\ \frac{\partial L(\boldsymbol{w},w_0,\boldsymbol{\xi}^-,\boldsymbol{\xi}^+,\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma})}{\partial \boldsymbol{w}_0} &= \sum_{i=1}^{\ell}\alpha_i y_i = 0. \end{aligned}$$

위의 정상(stationary) 조건들을 라그랑쥬 함수에 대입하고, 커널(kernel)을 사용하면, 아래의 쌍대 최적화 문제를 얻을 수 있다.

$$\begin{aligned} \underset{\boldsymbol{\alpha}}{\text{maximize}} \quad & \sum_{i=1}^{\ell}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) \\ \text{subject to} \quad & \sum_{i=1}^{\ell} y_i\alpha_i = 0 \geq 1-\xi_i^- + \xi_i^+, \\ & C_2 \leq \alpha_i \leq C_1,\ i=1,\ldots,\ell, \end{aligned} \tag{T_D}$$

여기서 $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \boldsymbol{\Phi}(\boldsymbol{x}_i) \cdot \boldsymbol{\Phi}(\boldsymbol{x}_j) \rangle$이고, $\boldsymbol{\Phi}$는 원래의 입력공간에서 고차원의 특성(feature)공간으로의 함수이다. 전형적인 커널함수들로서는, 가우지안(Gaussian) 커널함수와 p차의 다항(polynomial) 커널함수가 있다. 이 커널 함수들을 표현하면 아래와 같다:

$$K(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|^2}{\sigma^2}\right),$$
$$K(\boldsymbol{x}, \boldsymbol{y}) = (\langle \boldsymbol{x} \cdot \boldsymbol{y} \rangle + 1)^p.$$

본 논문의 수치 예제에서는 위에서 소개한 가우지안(Gaussian) 커널함수를 사용하여 실험을 실행한다.

이제, 바이어스(bias) w_0를 계산하기 위하여 α^*를 쌍대 문제 (T_D)의 최적해라 하자. 그리고 n_+를 $C_2 < \alpha_j^* < C_1$의 조건을 만족시키고 $y_j = 1$인의 $\boldsymbol{x}_j$개수라 하고, 또한 n_-를 $C_2 < \alpha_j^* < C_1$의 조건을 만족시키고 $y_j = -1$인의 $\boldsymbol{x}_j$개수라 하자. Karush-Kuhn-Tucker 보충(complementary)조건으로부터 만약 $C_2 < \alpha_i^* < C_1$이면, $\beta_i > 0$이고 $\gamma_i > 0$이다. 이것은 $\xi_i^- = \xi_i^+ = 0$임을 나타낸다. 따라서,

$$w_0^* = \frac{1}{n_- + n_+}\left((n_+ - n_-) - \sum_{j=1}^{n_+ + n_-}\sum_{i=1}^{\ell} y_j \alpha_i^* K(\boldsymbol{x}_i, \boldsymbol{x}_j)\right)$$

이다. 만약, $n_+ = n_- = 0$이면 당연히 $w_0 = 0$이다.

다음 절에서 다양한 C_1과 C_2값들에 대하여 기존의 soft margin 알고리즘과 새롭게 제안한 total margin 알고리즘을 비교한다. 아래에 소개하는 쌍대 ℓ_1-놈 soft margin 알고리즘 (S_D)문제는 1절에 소개된 soft margin 알고리즘의 1차 문제에 커널을 사용하여 표현하였다. 우리는 이것으로 두 알고리즘의 성능을 비교할 것이다.

$$\begin{aligned} \underset{\boldsymbol{\alpha}}{\text{maximize}} \quad & \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ \text{subject to} \quad & \sum_{i=1}^{\ell} y_i \alpha_i = 0, \\ & 0 \le \alpha_i \le C, \ i = 1, \ldots, \ell, \end{aligned} \tag{S_D}$$

4. 수치 예제들

이 절에서 기존의 soft margin 알고리즘과 total margin 알고리즘을 비교하는 수치 예제들의 결과들을 나타낸다. 본 실험에서는 가우지안 커널함수를 사용하여 여러 가지 경우의 벌칙 모수 값들에 대하여 수행하였다.

첫 번째 데이터는 BUPA Medical Research Ltd.로부터 얻어진 간장 장애(liver-disorder) 자료이다(http://www.ics.uci.edu/ mlearn/MLSummary.html).

이 데이터는 11개의 독립 변수와 345명의 남성 환자들의 관측값들로 구성이 되어 있다. 두 번째 데이터는 Cleveland Clinic Foundation에서 얻어진 심장 질환에 관한 자료이다. 이

표 4.1: 간장 질환 자료를 이용한 Soft Margin과 Total Margin에서 가우지안 커널 함수를 이용한 오분류율

C_1의 값	가우지안 커널		
	Soft Margin	Total Margin	
$C_1 = 1.0$	34.62	$C_2 = 0.05$	34.19
		$C_2 = 0.10$	**33.87**
		$C_2 = 0.50$	34.40
$C_1 = 5.0$	36.54	$C_2 = 0.10$	36.43
		$C_2 = 0.50$	35.36
		$C_2 = 1.00$	**34.40**
$C_1 = 10.0$	36.32	$C_2 = 0.10$	36.22
		$C_2 = 0.50$	36.00
		$C_2 = 1.00$	**34.94**

표 4.2: 심장병 자료를 이용한 Soft Margin과 Total Margin에서 가우지안 커널 함수를 이용한 오분류율

C_1의 값	가우지안 커널		
	Soft Margin	Total Margin	
$C_1 = 1.0$	20.00	$C_2 = 0.05$	20.00
		$C_2 = 0.10$	19.56
		$C_2 = 0.50$	**19.23**
$C_1 = 5.0$	21.21	$C_2 = 0.10$	21.10
		$C_2 = 0.50$	**20.22**
		$C_2 = 1.00$	20.99
$C_1 = 10.0$	21.32	$C_2 = 0.10$	21.32
		$C_2 = 0.50$	**19.87**
		$C_2 = 1.00$	21.10

표 4.3: Pima Indian 자료를 이용한 Soft Margin과 Total Margin에서 가우지안 커널 함수를 이용한 오분류율

C_1의 값	가우지안 커널		
	Soft Margin	Total Margin	
$C_1 = 1.0$	25.04	$C_2 = 0.05$	25.34
		$C_2 = 0.10$	25.41
		$C_2 = 0.50$	**24.35**
$C_1 = 5.0$	28.19	$C_2 = 0.10$	28.34
		$C_2 = 0.50$	**27.65**
		$C_2 = 1.00$	29.96
$C_1 = 10.0$	29.33	$C_2 = 0.10$	29.41
		$C_2 = 0.50$	**28.78**
		$C_2 = 1.00$	28.47

자료는 13개의 수치 값의 독립 변수와 303명의 환자에 대한 관측값이다. 세 번째 자료 집합은 당뇨병의 원인으로 생각되는 8개의 변수들과 환자들이 세계보건기구에서 정한 기준에 의해 당뇨병의 징후를 보이는가의 여부에 따라 두 집단으로 나눈 변수로 구성되어 있다. 이 자료에서 모든 환자들은 Pima Indian족의 21세 이상의 768명의 여성들을 대상으로 관측된 자료이다. 우리는 실험을 위하여 데이터 집합들에서 각각 70%와 30%로 훈련집합과 검사집합(testing set)으로 나누었다. 100번의 실험을 수행하여 얻어진 오분류율의 평균을 결과의 표에 나타내었다. 그 표들은 위에서 나타낸 바와 같다. 수치 예제들에서 기존의 soft margin 알고리즘과 새롭게 제안된 total margin 알고리즘의 오분류율의 값을 이용하여 비교하였다. Total margin 알고리즘을 이용하여 얻어진 결과들은 거의 대부분의 경우에서 soft margin 알고리즘을 이용한 경우보다 오분류율의 값이 낮게 나타난다. 표 4.1, 표 4.2 그리고 표 4.3들은 soft margin 분류 모형보다 total margin 알고리즘에서 분류율이 더욱 개선되었음을 확인할 수가 있다.

이러한 실험들을 통하여, 새롭게 제안된 방법이 기존의 soft margin 알고리즘을 이용한 분류보다 개선된 방법임을 알 수 있었다. 따라서, 우리는 soft margin 알고리즘보다 정분류된 점들도 함께 고려하는 total margin 알고리즘이 실험 결과들을 통하여 알 수 있는 바와 같이, 분류율에 있어서 효과적이라고 판단할 수 있다.

5. 결론

본 논문에서, 우리는 모든 데이터 점들과 분리 초평면 사이의 거리를 고려하는 새로운 total margin 알고리즘을 제안하였다. 그리고 선형 분류기들에 대한 일반화 오차 경계를 보이고 증명하였다. 이러한 접근 방법은 단지 soft margin 알고리즘의 slack 변수들뿐만 아니

라 정분류된 데이터 점들 사이의 거리의 측도인 surplus 변수들도 동시에 고려하고 있다. 이러한 새로운 알고리즘을 사용하여 기존의 soft margin 알고리즘이 개선됨을 분류율을 통하여 알 수 있었다. 또한, 이 방법은 모든 데이터 점들을 고려함으로써 기존의 서포트 벡터 기계 알고리즘들을 확장한 것임을 알 수 있다.

수치 예제들을 통해서, 다양한 C_1과 C_2값들에 대하여 기존의 soft margin 알고리즘과 새롭게 제안한 total margin 알고리즘을 비교할 때, 전반적으로 total margin 알고리즘이 soft margin 알고리즘보다 우수한 분류율을 보였다. 또한 total margin 알고리즘에 대한 일반화 오차 경계는 slack 변수들의 최소화와 surplus 변수들의 최대화에 의하여 감소됨을 알 수 있었다.

참고문헌

[1] 김철응, 윤민 (2003). 서포트 벡터 기계에서 잡음 영향의 효과적 조절, *응용통계연구*, Vol.16(2), 261-271.

[2] Alon, N., Ben-David, S., Cesa-Bianchi, N., and Hasssler, D. (1997). Scale-Sensitive Dimensions, Uniform Convergence, and Learnability. *Journal of ACM* , Vol.44, 615-631.

[3] 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.

[4] Bertsekas, D.P. (1995). *Nonlinear Programming*, Belmont, MA, Athena Scientific.

[5] Cherkassky, V. and Mulier, F. (1998). *Learning from Data Concepts, Theory, and Methods*, John Wiley & Sons, INC., New York.

[6] Cortes, C. (1995). *Prediction of Generalization Ability in Learning Systems*, Ph.D.Thesis, University of Rochester

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

[8] Freund, Y. and Schapire, R.E. (1998). Large margin classification using the perceptron algorithm, In Shavlik, J. editor *Machine Learning: Proceedings of the 15th International Conference*, Morgan Kaufmann.

[9] Freed, N. and Glover, F. (1981). Simple but Powerful Goal Programming Models for Discriminant Problems, *European Journal of Operational Research*, Vol.7, 44-60.

[10] Gurvits, L. (1997). A Note on a Scale-Sensitive Dimension of Linear Bounded Functionals in Banach Spaces, *In Proceedings of Algorithmic Learning Theory*, ALT-97.

[11] Haykin, S. (1998). *Neural Networks A Comprehensive Foundation.*(2nd ed.), Prentice Hall, New Jersey.

[12] 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.

[13] Schapire, R., Freund, Bartlett, Y. P. and Sun Lee, W. (1998). Boosting the Margin: A New Explanation for the Effectiveness of Voting methods, *Annals Statistics*, Vol.26, 1651-1686.

[14] Shawe-Taylor, J., Bartlett, P.L., Williamson, R.C., and Anthony, M. (1998). Structural Risk Minimization over Data-Dependent Hierarchies, *IEEE Transactions on Information Theory*, Vol.44, 1926-1940.

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

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

[17] Vapnik, V. (1999). *The Nature of Statistical Learning Theory.*(2nd ed.), Springer-Verlag, New York.

[2003년 2월 접수, 2003년 9월 채택]

Improving the Generalization Error Bound using Total margin in Support Vector Machines

Min Yoon [1)]

ABSTRACT

The Support Vector Machine(SVM) algorithm has paid attention on maximizing the shortest distance between sample points and discrimination hyperplane. This paper suggests the total margin algorithm which considers the distance between all data points and the separating hyperplane. The method extends existing support vector machine algorithm. In addition, this newly proposed method improves the generalization error bound. Numerical experiments show that the total margin algorithm provides good performance, comparing with the previous methods.

Keywords: Support Vector Machines(SVMs), Generalization Error Bound, Soft Margin, Surplus Variables, Total Margin

1) Part-time lecturer, Dept. of Applied Statistics, Yonsei University. E-mail: myoon@yonsei.ac.kr