Support Vector Machine의 입력데이터 오류에 대한 Robustness 분석

이상근[0] 장병탁
서울대학교 컴퓨터공학부
sklee[0]@bi.snu.ac.kr btzhang@cse.snu.ac.kr

Robustness Analysis of Support Vector Machines against Errors in Input Data

Sang-Kyun Lee[0] and Byoung-Tak Zhang
School of Computer Science and Engineering, Seoul National University

요 약

Support vector machine(SVM)은 최근 각광받는 기계학습 방법 중 하나로서, kernel function 이라는 사상(mapping)을 이용하여 입력 공간의 벡터를 classification 이 용이한 특징 (feature) 공간의 벡터로 변환하는 것을 근간으로 한다. SVM 은 이러한 특징 공간에서 두 클래스를 구분 짓는 hyperplane 을 일련의 최적화 방법론을 사용하여 찾아내며, 주어진 문제가 convex problem 인 경우 항상 global optimal solution 을 보장하는 등의 장점을 지닌다. 한편 bioinformatics 연구에서 주로 사용되는 데이터는 측정 오류 등 일련의 오류를 포함하고 있으며, 이러한 오류는 기계학습 방법론이 어떤 decision boundary 를 찾아내는가에 영향을 끼치게 된다. 특히 SVM 의 경우 이러한 오류는 특징 공간 벡터간의 관계를 나타내는 Gram matrix 를 변화로 나타나게 된다. 본 연구에서는 입력 공간에 오류가 발생할 때 그것이 SVM 의 decision boundary 를 어떻게 변화시키는가를 대표적인 두 가지 kernel function, 즉 linear kernel 과 Gaussian kernel 에 대해 분석하였다. Wisconsin 대학의 유방암(breast cancer) 데이터에 대해 실험한 결과, 데이터의 오류에 따른 SVM 의 classification 성능 변화 양상을 관찰하여 커널의 종류에 따라 SVM 이 어떠한 특성을 보이는가를 밝혀낼 수 있었다. 또 흥미롭게도 어떤 조건 하에서는 오류가 크더라도 오히려 SVM 의 성능이 향상되는 것을 발견했는데, 이것은 바꾸어 생각하면 Gram matrix 의 일부를 변경하여 SVM 의 성능 향상을 꾀할 수 있음을 나타낸다.

1. 서 론

일반적으로 우리가 어떤 데이터에 기계학습 방법론을 적용하여 classification 문제를 풀 때는 (1) 그 대상이 되는 데이터가 신뢰 가능하며, (2) 따라서 그 데이터로부터 학습된 분류에 대한 지식, 즉 decision boundary 역시 신뢰 가능하다는 전제가 깔려 있다. 이러한 전제 하에 학습에 사용되지 않은 새로운 데이터가 주어질 경우 학습된 decision boundary에 근거하여 새 데이터의 클래스를 결정하게 된다. 일반적으로 언급하는 일반화 성능 (generalization performance) 역시 위의 두 조건을 전제로 한다.

하지만 근래의 실험적 연구 분야, 특히 bioinformatics 분야에서 얻어지는 데이터는 실험 장비 등의 개선에도 불구하고 그 획득 과정의 복잡성으로 인해 여러 과정, 특히 측정 과정이나 전처리 과정에서 오류가 발생할 가능성을 내포하고 있다. 이러한 데이터를 기계학습에 이용할 경우, 기계학습 방법론이 찾아내는 decision boundary의 위치가 오류의 정도에 따라 달라질 수 있고 따라서 일반화 성능에도 악영향을 줄 수 있다. 그러므로 오류를 포함한 데이터에 기계학습 방법론을 적용하기 위해서는 우선 사용할 방법이 오류에 대해 어느 정도 민감한가를 분석하는 robustness analysis가 선행되어야 한다.

근래에 많이 연구되고 있는 기계학습 기법인 kernel method와 그에 기반한 support vector machine(SVM)은 다양한 종류의 데이터에 대해 기존 기계학습 방법론들에 비해 나은 성능을 보여주고 있다[1,3,6]. 하지만 실제로 kernel matrix를 구성할 때 주어져야 하는 kernel function의 parameter에 따라 그 성능의 변동이 심한 것 역시 사실이다. 이 문제에 대해 여러 가지 논의가 있었지만, robustness관점에서의 분석은 미흡한 실정이다.

본 논문에서는 입력 데이터에 다양한 정도의 오류를 발생시킬 때 classification 성능 변화를 측정하여 오류에 대한 SVM decision boundary의 robustness를 측정 및 분석하였다. 실험에는 Wisconsin 대학에서 제공되는 유방암 (breast cancer) 환자 분류 데이터[4]를 사용하였다. 이 데이터는 699명의 환자에 대한 9가지의 attribute 값을 포함하고 있으며, 비선형적인 decision boundary를 갖는 특징이 있어 기계학습 방법론의 성능을 측정할 때 자주 이용되는 데이터이다.

2. Kernel Robustness Analysis

2.1 Kernel Method

d차원의 입력 공간 데이터를 $x_1, x_2, ..., x_n$ 이라 하자. 입력 공간 X에서 특징 공간 Z으로의 변환 함수를 $\phi: X \rightarrow Z$ 라고 하면, Gram matrix K는 다음의 식으로 정의된다 [5].

$$K=\begin{bmatrix}\phi(x_1)^T\phi(x_1) & \phi(x_1^T)\phi(x_2) & \mathrm{L}\\ \phi(x_2^T)\phi(x_1) & \mathrm{O} & \\ \mathrm{M} & & \end{bmatrix}=Z^TZ$$

$$\text{where } Z=\begin{bmatrix}\phi(x_1^T)\\ \mathrm{M}\\ \phi(x_n^T)\end{bmatrix}$$

즉, K는 특징 공간 벡터의 모든 pairwise inner product를 담고 있다. K로 표현되는 선형사상을 $K(x,y)=<\phi(x),\phi(y)>$로 나타낼 수 있으며, 이것을 커널(kernel)이라고 한다. Mercer의 Theorem에 따르면 Gram matrix가 positive semi-definite (psd)이면 kernel function이 feature 공간에서의 inner product로 표현되도록 하는 함수 ϕ가 항상 존재하므로, psd인 Gram matrix를 만들어 낼 수만 있으면 ϕ를 explicit하게 정의하지 않더라도 원하는 결과를 얻을 수 있다. Gaussian kernel, spectrum kernel 등이 바로 이러한 형태의 커널들이다.

본 연구에서는 inner product 형태인 linear kernel과 radial 함수 형태인 Gaussian kernel을 연구 대상으로 하였다.

2.2 Support Vector Machine

Support vector machine(SVM)은 특징 공간에서 두 클래스의 데이터를 양분하는 hyperplane을 찾는 기법으로, hyperplane 근처의 데이터 포인트와 hyperplane과의 수직거리를 최대화하는 수식을 사용하여 hyperplane을 찾는다[5].

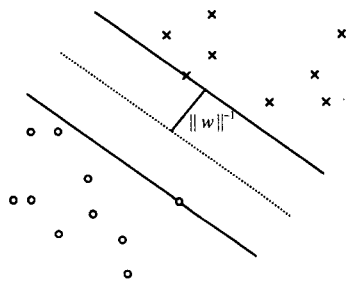

그림 1. SVM finds the separating plane maximizing the margin(length=$\|w\|^{-1}$)

클래스 레이블을 제외한 각 데이터를 x_i, 그 클래스 레이블을 $y_i\in\{+1,-1\}$라고 하면, hyperplan을 찾기 위한 수식은 다음과 같다.

$$\begin{cases}\min\ w^Tw\\ \text{s.t. } y_i(w^Tx_i+b)\geq 1\end{cases}$$

이것의 dual form은 다음과 같다.

$$\begin{cases}\max\ \theta(\alpha)=\sum_i\alpha_i-\frac{1}{2}\sum_{i,j}\alpha_i\alpha_jy_iy_jx_i^Tx_j\\ \text{s.t. } \sum_i\alpha_iy_i=0\end{cases}$$

이 형태의 문제는 quadratic programming을 이용하여 쉽게 해결할 수 있다. 또한, 목적함수와 최적해의 조건식이 모두 $<x_i,x_j>$의 inner product 형태로 나타나므로 Gram matrix를 정의함으로써 다양한 문제에 SVM을 적용할 수 있다.

3. 실 험

3.1 실험 설계

실험에는 미국 Wisconsin 대학의 유방암(breast cancer) 데이터를 사용하였다. 이 데이터는 총 699명의 환자에 대한 9가지 attribute 측정값을 포함하고 있으며, 두 클래스, 즉 양성 458명 (65.5%), 음성 241명 (34.5%)으로 구성되어 있다. 본 연구에서는 missing value가 있는 환자의 데이터를 제외하여 총 683명(전체 중 97.7%)의 데이터를 실험에 사용하였다. 실험에는 총 두 가지의 커널, 즉 linear kernel과 Gaussian kernel을 사용했다. 각 커널은 다음과 같이 표현된다.

Kernel	Equation
Linear kernel	$K(x,y)=<x,y>$
Gaussian kernel	$K(x,y)=\exp(-\frac{1}{\sigma^2}\|x-y\|^2)$

오류 ε는 입력 공간의 데이터에 발생할 수 있다고 가정했고, 정규분포 $N(0,s^2)$를 따르는 것으로 모델링하였다. 즉 오류가 있는 입력 공간은 $X'=X+\varepsilon$로 기술된다. 이 때 위의 커널 들은 다음과 같이 변경된다.

Kernel	Equation
Linear kernel	$K(x',y')=<x+\varepsilon,y+\varepsilon'>$
Gaussian kernel	$K(x',y')=\exp(-\frac{\|x-y+\varepsilon-\varepsilon'\|^2}{\sigma^2})$

그 다음 오류 ε의 확률분포, 즉 $N(0,s^2)$의 s값을 0~10으로 변화시키면서 SVM의 classification 성능 변화를 관찰하였다. Linear classifier의 경우 regularization parameter의 값에 따라 큰 성능 변화는 없었으므로 c=10의 값을 사용했다. 또한 Gaussian kernel의 경우는 σ의 값을 [1,10] 사이에서 0.5 단위로 변화시키면서 모든 패턴을 올바로 분류해 내는 σ=7.5 값을 찾아냈다. 여기서 parameter값이 [0,10]의 값을 갖는 것은 사용한 데이터의 attribute 값이 [0,10] 사이의 값을 갖기 때문이다.

3.2 실험 결과

3.2.1 Linear Kernel

Linear kernel의 경우, 오류가 없을 때 (s=0) 전체 683명 중 40명을 잘못 분류했다 (오차율 5.86%). 한편 s값의 변화에 따른 classification 성능 변화는 그림 2와 같다 (각 noise level에 대해 100회씩 반복 실험하였다).

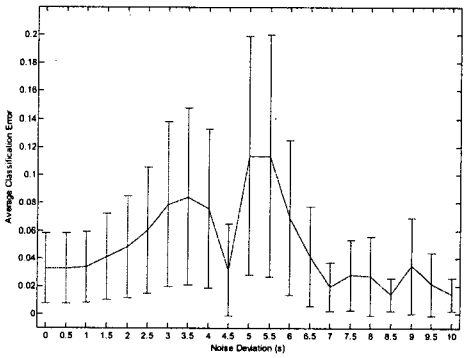

그림 2. The change of classification performance (linear kernel)

여기서 관찰할 수 있는 사실은 (1) s값이 0~6 정도에서는 예측할 수 있는 바와 같이 오류가 증가할수록 classification error 역시 증가한다는 것이다. 한편, (2) s>6 에서는 오히려 오류가 감소하는 것을 관찰할 수 있는데, 특히 s=7 또는 8.5에서는 오류가 없는(s=0) 경우보다도 classification error가 작아지는 현상을 발견하였다.

3.2.2 Gaussian Kernel

Gaussian kernel의 경우는 linear kernel의 경우와는 달리 오류가 증가해도 classification 성능은 크게 변하지 않았다 (그림 3). 그 이유는 kernel function을 사용하여 설명 가능하다. 즉 linear kernel의 수식은,

$$K(x',y') = < x+\varepsilon, y+\varepsilon' >$$
$$= K(x,y) + < \varepsilon, y > + < x, \varepsilon > + \| \varepsilon \|^2$$

인 반면, Gaussian kernel의 경우는,

$$K(x',y') = \exp(-\frac{1}{\sigma^2}\| x-y+\varepsilon-\varepsilon' \|^2)$$
$$= \exp\left\{-\frac{1}{\sigma^2}\left(\| x-y \|^2 + 2\| x-y \| \cdot \| \varepsilon-\varepsilon' \| + \| \varepsilon-\varepsilon' \|^2\right)\right\}$$

이 된다. 즉, linear kernel의 경우는 오류의 norm이 원래의 함수 값에 더해지는 형태인 반면, Gaussian kernel의 경우는 두 데이터 포인트가 가진 오류값의 차의 norm이 더해지므로, Gaussian kernel 오류항의 영향이 linear kernel의 그것보다 적을 수 있는 것이다.

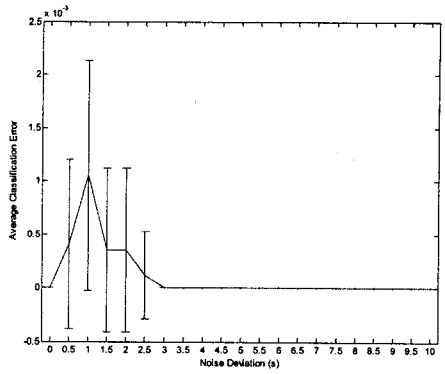

그림 3. The change of classification performance (Gaussian kernel)

4. 결 론

본 논문에서는 데이터에 발생할 수 있는 오류에 대한 간단한 확률모델을 가정하고, linear kernel과 Gaussian kernel을 사용할 경우에 SVM classifier가 오류에 대해 어떠한 반응을 보이는가를 측정하고 분석하였다. 그 결과 radial basis function(RBF)의 형태를 갖는 Gaussian kernel이 보다 더 오류에 robust함을 알 수 있었다. 또한 그 원인이 $f(x-y)$의 형태로 표현되는 RBF의 특성에서 기인함을 보였는데, 이것은 RBF 형태의 모든 커널에 적용되는 결과이다.

또한 두 커널 모두에서 오류율이 큰 경우 오히려 SVM classifier의 성능이 개선되는 경우가 있었는데, 이것은 역으로 해석하면 kernel의 변경 없이도 Gram matrix를 일부 수정하여 SVM classifier의 성능을 개선할 수 있음을 간접 증명하는 것이다. 이 사실은 SVM 성능 개선을 위한 후속 연구를 위한 기초 자료로서 중요한 역할을 할 것으로 기대된다.

감사의 글

본 연구는 과학기술부 국가지정연구실(NRL) 사업에 의해 지원되었음.

참고문헌

[1] Ben-Hur A., Horn D., Siegelmann H.T., Vapnik V.: A Support Vector Method for Clustering. *Neural Information Processing Systems* 2000: 367-373.
[2] Duda R.O., Hart P.E., Stork D.G., Pattern Classification, 2nd Ed, Wiley-Interscience, 2000.
[3] Guyon I., Weston J., Barnhill S., Vapnik V.: Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(1-3): 389-422, 2002.
[4] Mangasarian O.L., Wolberg W.H., Cancer diagnosis via linear programming, *SIAM News*, 23:5, pp1,18, 1990.
[5] Scholkopf B., Smola A.J, Learning with Kernels: Support vector Machines, Regularization, Optimization, and Beyond, MIT Press, 2001
[6] Scholkopf B., Tsuda K., Vert J-P., Kernel Methods in Computational Biology, MIT Press, 2004