신경 회로망 학습을 통한 모델 선택의 자동화

Automation of Model Selection through Neural Networks Learning

류 재 홍
여수대학교 컴퓨터공학과

Jae Hung Yoo
Dept. of Computer Engineering, Yosu National University
E-mail : jhy@ce.yosu.ac.kr

Abstract

Model selection is the process that sets up the regularization parameter in the support vector machine or regularization network by using the external methods such as general cross validation or L-curve criterion. This paper suggests that the regularization parameter can be obtained simultaneously within the learning process of neural networks without resort to separate selection methods. In this paper, extended kernel method is introduced. The relationship between regularization parameter and the bias term in the extended kernel is established. Experimental results show the effectiveness of the new model selection method.

1. 서론

일차 연립 방정식에서 일반적인 경우 문제의 해를 구하기 위해 역행렬(inverse) 또는 유사 역행렬(pseudo inverse)을 취하여 간단히 해를 구할 수 있지만 불량 조건 문제(ill-conditioned problem) 또는 특이 시스템(singular system)의 경우 역 행렬을 취하게 되면 그 값이 너무 커지거나 무한하게 된다. 이 경우 오차와 해의 크기가 동시에 작아야 된다는 조건을 세워 불량 조건을 양호 조건(well-conditioned)으로 특이 시스템을 정상 시스템(usual system)으로 만드는 것을 정규화(regularization)라 한다[7].

모델 선택(model selection)은 SVM(Support Vector Machine)[1]과 RN(Regularization Networks)[3] 등 커널 방법(Kernel Methods)에서 가중치 학습과 분리된 자유 인수(free parameter)로 간주되는 티코노브 정규화 인수(Tikhonov regularization Parameter)를 구하는 절차다. 노이즈 수준을 알 경우는 분산 원칙(discrepancy principle)이 사용되고 노이즈 수준을 알 수 없는 경우는 GCV(General Cross Validation)와 L-곡선 검정(L-curve criterion) 이 대표적인 모델 선택 방법이다[4, 6].

본 논문은 기존에 별도로 구하던 정규화 인수를 신경망 학습에 포함해서 동시에 구하는 방법을 제시한다. 바이어스(bias) 또는 문턱(threshold) 항을 포함한 확장된 커널 구조(extended kernel architecture)를 소개하고 기존 커널의 정규화 인수와 확장된 커널의 바이어스 항과의 관계를 설정한다. 확장된 커널은 커널 선형 판별식(KLDA Kernel Linear Discriminant Analysis) 학습 방법들로 가중치를 구한다.

본 논문의 구성은 다음과 같다. 2장에서는 정규화와 모델 선택에 대한 기존의 방법들을 기술한다. 3장에서는 제안하는 확장된 커널 시스템에 의한 정규화 효과에 대하여 논의하고 4장에서 확장된 커널시스템에 대한 학습법에 대하여 논한다. 5장은 실험 결과이고 6장은 결론이다.

2. 정규화와 모델선택

불량자세 문제(ill-posed problem)는 Hadamard가 1902년 편미분 방정식에 관한 논문에서 소개한 개념으로 다음의 세 가지 조건 중 하나 이상 만족하면 불량자세이다[5].

1. 해가 존재하지 않는다.
2. 해가 유일하지 않다.
3. 해가 자료에 연속적으로 종속이지 않다.

일차 연립 방정식에서 과결정(over-determined) 시스템은 1번 조건을 만족하며 급결정(under-determined) 시스템은 2번 조건을 만족한다. 급결정 또는 과결정 시스템에서 행렬이 정상 조건(well-conditioned)이면 유사 역 행렬에 의하여 최소 길이 해 또는 근사 해를 구할 수 있다. 선형 시스템이 불량 조건(ill-conditioned) 이면 즉 행렬의 최대 최소 특성치 비율(ratio of max/min eigenvalues)로 정의되는 조건 수(condition number)가 아주 크거나 무한대인 경우의 3번 조건을 만족한다[8]. 3번 조건을 다시 설명하면 해가 자료의 섭동(perturbations on data)에 불안정(unstable)한 것이다. 즉 평활 조건(smoothness condition)을 만족하지 못하는 것이다.

Tikhonov 정규화는 불량자세 문제 중에서 3번째 조건을 만족하는 시스템에서 출력 오차와 해에 대한 평활 연산자를 갖는 비용 함수(cost function)를 최적화하는 해를 구하는 문제다. 선형시스템에서 평활 연산자가 단위 행렬(identity matrix)일 경우 다음과 같다[4,6,9].

$$C(\mathbf{a},\lambda)=\frac{1}{2}\|\boldsymbol{\varepsilon}\|^2+\lambda\frac{1}{2}\|\mathbf{a}\|^2 \quad (2.1)$$
$$\boldsymbol{\varepsilon} = \mathbf{d}-\mathbf{Y}\,\mathbf{a}$$

커널 방법에서는 다음과 같다[3,7].

$$C(\boldsymbol{\alpha},\lambda)=\frac{1}{2}\|\boldsymbol{\varepsilon}\|^2+\lambda\frac{1}{2}\,\boldsymbol{\alpha}^t\mathbf{K}\,\boldsymbol{\alpha}, \quad (2.2)$$
$$\boldsymbol{\varepsilon} = \mathbf{d}-\mathbf{K}\,\boldsymbol{\alpha}$$

위 식을 가중치 $\boldsymbol{\alpha}$에 대하여 미분하여 영으로 놓으면 다음 식을 얻는다.

$$(\mathbf{K}+\lambda\,\mathbf{I})\,\boldsymbol{\alpha} = \mathbf{d} \quad (2.3)$$

주어진 Tikhonov 정규화 인수 λ는 행렬 $(\mathbf{K}+\lambda\mathbf{I})$가 정상 조건(well-conditioned)이 되는 스칼라(scalar) 값을 가진다. 따라서 위 식은 아래와 같이 $\boldsymbol{\alpha}$의 해를 구할 수 있다.

$$\boldsymbol{\alpha}_{reg} = (\mathbf{K}+\lambda\,\mathbf{I})^{-1}\mathbf{d} \quad (2.4)$$

정규화 인수는 식(2.4)를 식(2.2)에 대입하여 구할 수 없다. 노이즈 수준 δ를 알 경우는 분산 원칙(Discrepancy Principle)이 사용하여 다음 식을 푼다.

$$\|\mathbf{d}-\mathbf{K}\boldsymbol{\alpha}_{reg}\| = \delta \quad (2.5)$$

노이즈 수준을 알 수 없는 경우는 GCV(General Cross Validation)와 L-곡선 검정(L-curve Criterion) 이 대표적인 모델 선택 방법이다. GCV에서 최적의 정규화 인수 λ는 주어진 자료에 대해 아래의 목적 함수를 최소화한다[4, 7].

$$V(\lambda)=\frac{\frac{1}{N}\|(\mathbf{I}-\mathbf{A}(\lambda))\,\mathbf{d}\|^2}{\left[\frac{1}{N}tr[\mathbf{I}-\mathbf{A}(\lambda)]\right]^2} \quad (2.6)$$

여기서 $\mathbf{A}(\lambda)=\mathbf{K}(\mathbf{K}+\lambda\mathbf{I})^{-1}$ 이고 N은 자료 개수이다.

Tikhonov 비용함수를 나타내는 식(2.2)과 식(2.4)를 참조하면 정규화 인수 λ의 변화에 따라 변수 $\log\|\boldsymbol{\alpha}^t_{reg}\mathbf{K}\,\boldsymbol{\alpha}_{reg}\|_2^{1/2}$과 $\log\|\mathbf{d}-\mathbf{K}\boldsymbol{\alpha}_{reg}\|_2$를 종축과 횡축에 두는 그래프는 L자 모양의 곡선이 된다. 이때 최적화 정규화 인수는 수직 좌표의 섭동(perturbation)오차와 수평 좌표의 잔여오차(residual error)의 균형을 가장 잘 이룰 수 있는 점에서 정의된다. 이 점은 그림에서 수직과 수평 그래프가 이루는 코너에 존재한다. L-Curve 검정에서는 최적화 정규화 인수에 해당하는 점을 근사적으로 찾기 위해 코너에서 최대 곡률(curvature)을 가진 점을 계산한다[6].

3. 확장 커널 시스템 개발

2장에서 검토한 바 기존의 정규화 방법은 먼저 분산원칙, L-Curve 또는 GCV 등의 방법으로 최적의 정규화 인수를 찾은 후에 가중치를 계산하는 방법이다. 본 논문에서는 불량조건문제 또는 특이 커널에 대해 커널을 확장하여 가중치와 바이어스를 동시에 학습하게 함으로써 정규화 인수와 가중치의 학습을 하나로 통합한 새로운 정규화 방법을 제안한다.

바이어스 항을 포함하는 RBF 네트워크의 시스템은 아래와 같은 급결정 선형 시스템(under-determined linear system)이다[7].

$$\mathbf{K}\,\boldsymbol{\alpha} + \alpha_0 = \mathbf{d} \quad (3.1)$$

정규화 RBF 네트워크는 $\lambda\boldsymbol{\alpha}$ 항을 포함하면 다음과 같다.[7].

$$(\mathbf{K}+\lambda\,\mathbf{I})\,\boldsymbol{\alpha} = \mathbf{d} \quad (3.2)$$

릿지 회귀 분석(Ridge Regression)에서 입력 패턴 $\mathbf{x}_i$와 출력 d_i에 대해 아래와 같이 정의된다[7].

$$d_i = f(\mathbf{x}_i) + \varepsilon_i \qquad E[\varepsilon_i] = 0 \tag{3.3}$$

즉 오차벡터는 일반적으로 백색화가 된 것으로 가정한다(white noise assumption). 이때 식(3.1)과 식(3.2)의 바이어스 항 α_0 와 정규화 인수 λ의 관계를 구하면 아래와 같다.

$$\begin{aligned}\mathbf{d} &= \mathbf{K}\boldsymbol{\alpha} + \lambda\boldsymbol{\alpha} \\ &= \mathbf{K}\boldsymbol{\alpha} + \bar{\xi} \\ &= \mathbf{K}\boldsymbol{\alpha} + \bar{\xi} + (\xi - \bar{\xi}) \\ &= \mathbf{K}\boldsymbol{\alpha} + \alpha_0 + \boldsymbol{\varepsilon}\end{aligned} \tag{3.4}$$

바이어스 항 α_0 와 정규화 인수 λ, 가중치 $\boldsymbol{\alpha}$의 관계를 정립하면 다음과 같다.

$$\alpha_0 = \bar{\xi} \equiv \frac{\sum \xi_i}{N} = \frac{\lambda \sum \alpha_i}{N} \tag{3.5}$$

또는

$$\lambda = \frac{\alpha_0}{\sum \alpha_i} N \tag{3.6}$$

정규화 인수와 가중치 벡터의 곱은 오차 벡터가 된다. 하지만 이 오차벡터는 일반적으로 백색화가 된 것은 아니며 백색화를 위하여 오차의 평균을 구하면 '0' 이 아니다. 본 논문에서는 이 오차의 평균을 바이어스 항으로 해석할 것을 제안하는 것이다. 이것으로 또한 기존 정규화 방법에서 식 (3.2)의 $\lambda\boldsymbol{\alpha}$ 항을 가중치 계산에만 적용하고 출력 벡터 $\mathbf{d}$를 추정하는 데 더하지 않는 가에 대한 의문점을 해소하는 계기가 될 것이다. $\lambda\boldsymbol{\alpha}$ 항을 더하여 출력 벡터 $\mathbf{d}$를 추정하면 훈련 자료에 대한 과도한 짜 맞춤(over-fitting)이 되어 평가 자료에 대한 일반화는 불가능하다.

4. 확장 커널 학습 방법

커널 LMS는 기존의 선형판별식 학습법인 LMS를 커널 판별식에 적용한 것이다[2, 7, 10].

Table. 4.1 KLMS Algorithm

Given Training Data $\mathbf{Y}$, $\mathbf{d}$
Construct Kernel $\mathbf{K}$
Begin initialize $\boldsymbol{\alpha} = \mathbf{0}$, $\eta(\cdot)$,
margin b > 0, tolerance θ, $k = 0$
do
Shuffling the training data

$$\boldsymbol{\alpha}(k+1)' = \boldsymbol{\alpha}(k) + \eta(k)\frac{d_k b - \mathbf{K}(k,:)\,\boldsymbol{\alpha}}{\| \mathbf{K}(k,:) \|^2}\mathbf{K}(k,:)$$

until
$| \eta(k) (d_k b - \mathbf{K}(k,:)\,\boldsymbol{\alpha}) | < \theta$
return $\boldsymbol{\alpha}$
End

커널 학습 방법을 적용하기 위해서는 먼저 커널을 패턴 행렬은 바이어스 α_0를 학습하기 위해 아래와 같이 확장하였다.

$$\begin{aligned}&\mathbf{H}\,\boldsymbol{\omega} = \mathbf{d} \\ &\mathbf{H} \equiv [\,\mathbf{K} \quad ones(N,1)\,], \\ &\boldsymbol{\omega} \equiv \begin{bmatrix} \boldsymbol{\alpha} \\ \alpha_0 \end{bmatrix}\end{aligned} \tag{4.1}$$

이상과 같은 방법으로 바이어스 α_0는 패턴 가중치 $\boldsymbol{\alpha}$가 학습하는 동안 함께 학습함으로써 기존의 사용자 지정 방식의 정규화 인수는 패턴가중치와 함께 학습하게 된다. 이때 확장 커널 판별식 $\mathbf{H}\,\boldsymbol{\omega} = \mathbf{d}$는 선형 판별식 $\mathbf{Y}\,\boldsymbol{\alpha} = \mathbf{d}$와 대응한다.

5. 실험결과

표준편차의 원들이 상호간에 겹치는 가우시안 데이터는 영역이 겹침으로 비 선형 분류기로도 완전한 분리를 할 수 없다. 그림 Fig4.1에서 좌측의 큰 원은 최적 베이지안 결정 경계를 나타내며 우측의 원과 베이지안 결정 경계 안의 작은 원은 가우시안 데이터의 중점과 표준편차에 의해 그려진 원이다.

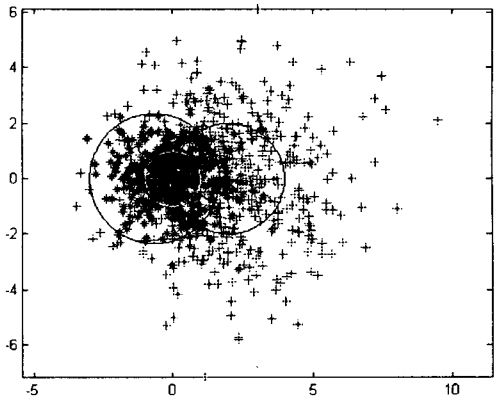

Fig. 5.1 One thousand testing patterns of Gaussian distribution

이때 사용한 평균 벡터(mean vector) $\boldsymbol{\mu}$와 분산(variance) $\boldsymbol{\sigma}$은 각각 아래와 같다[7].

Table. 5.1 Condition of Gaussian data

	Class1	Class2
$\boldsymbol{\mu}$	$[0\ 0]^T$	$[2\ 0]^T$
$\boldsymbol{\sigma}$	$[1\ 1]^T$	$[2\ 2]^T$

커널 LMS방법은 바이어스 포함여부에 따라 EKLMS와 KLMS로 명명하고 분류 성능을 비교하면 최적인 베이지안 분류기의 분류 성능이 81.51%이므로 기존의 정규화 방법과 확장 커널 LMS 방법 모두 동등한 최적의 분류 성능이 있

는 것으로 평가한다.

Table. 5.2 Classification Performance

Training Testing	Inverse	L-Curve	GCV	KLMS	EKLMS
Without Bias	52.8 51.4	82.2 82.5	82.4 82.6	78.8 78.3	81.8 80.4
With Bias	N/A	82.2 83.1	82.2 82.9	N/A	82.4 82.9

1차원 이미지 복구 문제인 Shaw 데이터를 이용하여 각각의 알고리듬의 성능을 비교해 보았다[6]. 실험결과 바이어스 항을 포함한 경우 오류가 낮아지는 것을 볼 수 있으며 전체적인 결과는 L-Curve, GCV, EKLMS, Inverse 순으로 나타났으나 역행렬을 제외한 각 알고리듬들 간의 오류는 매우 근소한 것으로 나타났다.

Table. 5.3 Estimation of The Regularization parameter using Shaw Data

	λ	Vector Norm	RMS Error	Cost	RMS Error with Bias	Cost with Bias
Inverse	0	1.920854 e+015	1.090466 e+007	1.902587 e+015	N/A	N/A
L-Curve	8.728774 e-004	8.207618 e+000	7.404211 e-004	2.940946 e-002	7.762955 e-004	2.941033 e-002
GCV	4.604075 e-003	5.904194 e+000	4.388925 e-003	8.055611 e-002	2.735985 e-003	8.036767 e-002
EKLMS	2.477042 e-002	4.493981 e+000	1.690067 e-002	3.808288 e-001	1.173942 e-002	3.784637 e-001

실험에 적용한 수식은 다음과 같다.

Table. 5.4 Equations for Experiment Analysis

- $\boldsymbol{\alpha}=(\mathbf{K}+\lambda\mathbf{I})^{-1}\,\mathbf{d}$
- $\alpha_0=\lambda\dfrac{\sum\alpha_i}{N}$
- Vector Norm = $\|\boldsymbol{\alpha}\|_2$
- RMS Error = $\sqrt{\dfrac{\|\mathbf{d}-\mathbf{K}\,\boldsymbol{\alpha}\|^2}{N}}$
- RMS Error with Bias = $\sqrt{\dfrac{\|\mathbf{d}-(\mathbf{K}\,\boldsymbol{\alpha}+\alpha_0)\|^2}{N}}$
- Cost = $\dfrac{\|\mathbf{d}-\mathbf{K}\,\boldsymbol{\alpha}\|^2}{2}+\lambda\dfrac{\boldsymbol{\alpha}'\,\mathbf{K}\,\boldsymbol{\alpha}}{2}$
- Cost with Bias = $\dfrac{\|\mathbf{d}-(\mathbf{K}\,\boldsymbol{\alpha}+\alpha_0)\|^2}{2}+\lambda\dfrac{\boldsymbol{\alpha}'\,\mathbf{K}\,\boldsymbol{\alpha}}{2}$

5. 결론

본 논문에서는 기존에 별도로 구하던 정규화 인수를 신경망 학습에 포함해서 동시에 구하는 방법을 제시하였다. 바이어스를 포함한 확장된 커널 시스템을 소개하고 정규화 인수와 바이어스 항과의 관계를 설정하였다. 확장된 커널은 커널 선형 판별식 학습 방법들에 의해서 가중치 학습을 수행한다. 시험결과 기존의 정규화 네트워크 RN과 동등한 수행 능력을 확인하였다. 향후 과제로는 저밀도 표현(sparse representation) 즉 주어진 데이터 수보다 작은 수의 지원 벡터 개수(number of support vectors) 또는 가중치 $\boldsymbol{\alpha}$의 비영 요소의 수(number of nonzero elements)를 갖는 커널 방법을 연구하는 것이다.

6. 참고문헌

[1] B. E. Boser, et al., "A training Algorithm for optimal margin classifiers," in *Proc. of the 5th Annual Workshop on Computational Learning Theory 5*, pp. 144-152, Pittsburgh, 1992.

[2] R. O. Duda et al., *Pattern Classification and Scene Analysis,* John Wiley & Sons, 1973.

[3] F. Girosi, et al., "Regularization theory and neural networks architectures," *Neural Computation,* Vol. 7, pp. 219-269, Pittsburgh, 1995.

[4] G. Golub, et al., "Generalized cross-validation as a method for choosing a good ridge parameter," *Techonometrics*, vol. 21 pp. 215-223, , 1979.

[5] J. Hadamard, "Sur les problemes aux derivees partielles et leur signification physique." *Princeton University Bulletin*, pp. 49-52, 1902. cited in http://en.wikipedia.org/wiki/Well-posed_problem

[6] P. C. Hansen, "Regularization Tools, A Matlab Package for Analysis and Solution of Discrete Ill-Posed Problems," Version 3.1 for Matlab 6.0, 2001. http://www.imm.dtu.dk/~pch/Regutools/regutools.html

[7] S. Haykin, *Neural Networks - A Comprehensive Foundation,* 2nd Ed., Prentice-Hall, 1999.

[8] D. G. Luenberger, *Linear and Nonlinear Programming,* 2nd Ed., Addison-Wesley, 1984.

[9] A. N. Tikhonov, et al., *Solutions of Ill-posed Problems* V H Winston and sons, Washington D.C. 1977.

[10] B. Widrow, et al., "Adaptive switching circuits." in *1960 IRE WESCON Convention Record,* pp. 96-104, IRE. New York, 1960.