패턴인식을 위한 오차함수의 최적해

Optimal Solutions for Various Error Functions

오 상 훈
목원대학교

Sang-Hoon Oh
Mokwon University

요약

패턴인식 문제의 학습을 위하여 여러 형태의 오차 함수들이 제안되었다. 이 논문에서는 이들 오차함수들에 대하여 그 특징을 통계학적으로 분석하여 비교하였다. 이 분석결과는 패턴인식기의 학습에 있어서 적합한 오차함수를 선정하는 이론적 토대를 마련해준다.

I. 서론

패턴인식은 상당히 다양한 응용 분야에 걸쳐 있으며, 이를 학습시키는 방법 및 모델 역시 다양하다. 패턴인식기는 상당히 다양한 방법으로 구성될 수 있는데, 가장 인기 있는 방법이 discriminant 함수를 사용하는 것이다. 이 경우, discriminant 함수 값은 인식에 대한 신뢰도를 나타낸다. 즉, 인식기의 판단은 가장 큰 discriminant 값을 지닌 클래스를 선택하는 것이다[1]. 클래스에 속하는 신뢰도를 나타내는 더 자연스러운 방법은 확률함수를 사용하는 것이다. 확률함수가 주어지면 Bayes 이론을 근거로 최적의 Bayes 인식기를 구현할 수 있다[2].

Bayes 인식기를 구현 할려면 패턴에 대한 확률밀도함수를 알아야 하는데, 확률밀도함수를 정확하게 알 수 있느냐 하는 것이 Bayes 인식기 구현의 성패를 좌우한다. 학습패턴으로부터 확률밀도함수를 예측하는 방법으로 nonparametric estimation이 있다. Parzen window 방법이 여기에 속하는데, 이 방법은 각 샘플의 위치에 창함수를 위치하도록 하여 패턴의 확률밀도함수를 예측한다[3]. 이러한 방법은 상당히 많은 수의 샘플과 아주 많은 kernel 함수들이 필요하다.

또 다른 방법으로는, discriminant 함수를 지닌 인식기를 구성하는 것인데, 이 경우 확률밀도함수 예측에서 사용하는 것보다 훨씬 더 적은 kernel 함수를 사용하게 되며, 인식기의 출력은 클래스에 속할 확률을 나타내지 않아도 된다. 또 대부분의 경우 discriminant 함수를 사용한 인식기가 확률밀도함수 예측을 기반으로 한 인식기보다 인식률의 관점에서 더 좋은 성능을 보이는 경우도 많다.

따라서, 대부분의 패턴 인식기들은 discriminant 함수를 출력하도록 학습이 되는데, 이 경우 출력이 원하는 값이 되도록 하기 위하여 학습에 필요한 오차(error) 함수를 정의하게 된다. 대표적인 오차함수는 MSE(mean-squared error)를 들 수 있다[4]. 또한 패턴인식기에 많이 응용되는 신경회로망의 학습을 보다 더 잘하기 위하여 CE(cross-entropy) 오차함수도 제안되었으며[5], 이의 성능을 더 개선한 nCE(*n*th order extension of CE) 오차함수도 있다[6]. 가장 최근에는 데이터 불균형이 심한 경우에 패턴인식기의 성능을 향상시키기 위하여 제안된 오차함수들도 있다[7].

이 논문에서는 이렇게 제안된 오차함수들의 통계적인 특성을 분석하고 비교하여, 패턴인식기를 구현하는 경우 오차함수를 선택하는 데 유용한 정보를 제공하고자 한다.

Ⅱ. 다양한 오차함수들의 통계학적 비교

패턴 인식기에 임의의 학습패턴이

$$\mathbf{x}=[x_1^{(p)},x_2^{(p)},\ldots,x_M^{(p)}]^T \tag{1}$$

이 주어졌다고 하자. 여기서, $p=1,2,\ldots,P$는 학습패턴의 인덱스이다. 이 패턴이 특정 클래스에 속하는 정보는 인식기의 출력 $\mathbf{y}=[y_1,y_2,\ldots,y_M]^T$ 에 대한 목표벡터에

$$t_k=\begin{cases}+1 \text{ if } \mathrm{x}\in C_k\\ -1\ otherwise\end{cases} \tag{2}$$

와 같이 주어진다고 가정하자. 여기서, C_k는 클래스 k의 집합을 나타낸다.

Discriminant 함수를 출력하는 패턴 인식기의 학습은 임의의 학습패턴이 입력될 때, 출력 벡터 $\mathbf{y}$와 목표벡터 $\mathbf{t}$ 사이의 오차가 최소가 되도록 하는 것이다. 이 오차를 보통 MSE 함수로 주어지는

$$E_m=\frac{1}{P}\sum_{p=1}^{P}\frac{1}{2}\sum_{k=1}^{M}(t_k^{(p)}-y_k^{(p)})^2 \tag{3}$$

를 사용한다[4]. $P\to\infty$ 일 때, 식 (3)을 최소화 시키는

것은

$$E\left\{\frac{1}{2}\sum_{k=1}^{M}(T_k - y_k(\mathbf{X}))^2\right\} \quad (4)$$

을 최소화 시키는 것이다[1]. 여기서, $E\{.\}$는 기대치를 나타내며, T_k는 목표값을 나타내는 확률변수이며, $\mathbf{X}$는 입력 패턴을 나타내는 확률벡터이다. 따라서, 식 (4)가 최소가 되는 해(solution) $b_k(\mathbf{x})$는

$$b_k(\mathbf{x}) = E\{T_k|\mathbf{x}\} \quad (5)$$

와 같이 구해진다. 한편, 목표벡터가 식 (2)와 같이 부호화 되었으므로,

$$E\{T_k|\mathbf{x}\} = 2Q_k(\mathbf{x}) - 1 \quad (6)$$

이 되며, $Q_k(\mathbf{x})$는 사전확률

$$Q_k(\mathbf{x}) = \Pr[\mathbf{X} \in C_k|\mathbf{X} = \mathbf{x}] \quad (7)$$

이다.

패턴인식기의 학습을 위하여 주어진 CE 오차함수는

$$E_{CE} = -\frac{1}{P}\sum_{p=1}^{P}\sum_{k=1}^{M}[(1+t_k^{(p)})\ln(1+y_k^{(p)}) + (1-t_k^{(p)})\ln(1-y_k^{(p)})] \quad (8)$$

로 주어진다[5]. 이 경우에도 식 (8)를 최소화 시키는 해는 식 (5)로 주어진다.

CE 오차함수를 확장한 nCE 오차함수[6]의 경우에도 위와 같은 절차에 의해 최적의 해를 유도할 수 있다. 그리고, 최근에 제안된 불균형 데이터의 경우 패턴인식기의 학습을 위하여 제안된 오차함수[7]에 대하여도 같은 절차를 적용하여 최적 출력 값의 해를 유도한다.

위와 같이 MSE, CE, nCE, 그리고 [7]에 주어진 오차함수의 최적해들을 비교해보면, MSE와 CE는 $b_k(\mathbf{x})$가 $Q_k(\mathbf{x})$와 1차함수의 관계를 보여주지만, nCE의 경우는 $b_k(\mathbf{x})$가 $Q_k(\mathbf{x})$와 고차함수의 관계를 지님을 볼 수 있다. 특히, nCE의 경우 n의 값을 조정하여 $b_k(\mathbf{x})$의 모양을 조절할 수 있음도 알 수 있다. 즉, n의 조절로 패턴인식기의 성능을 조절된다. 또한, 데이터 불균형의 경우를 다루는 패턴인식기의 학습을 위하여 제안된 [7]의 경우는 오차함수의 차수 파라미터를 조절하여, 출력값들 사아의 최적해 관계도 조절할 수 있다.

Ⅲ. 결론

이 논문에서는 패턴인식기의 학습을 위하여 제안된 여러 가지 오차함수들에 대하여 통계학적 분석을 통하여 얻어진 최적해들을 비교하였다. MSE와 CE의 경우는 통계학적으로 최적해가 같은 모양을 보인다. nCE 오차함수의 경우는 파라미터 조정을 통하여 최적해의 모양을 조절할 수 있다. 데이터 불균형의 경우에 대하여 제안된 오차함수[7]의 경우는 출력 값들 사이의 최적해 관계도 달리 할 수 있다.

■ 참 고 문 헌 ■

[1] H. White, "Learning in artificial neural networks: a statistical perspective," Neural Computa., vol. 1, pp. 425-464, 1989.

[2] K. Fukunaga and D. Kessell, "Nonparametric bayes error estimation using unclassified samples," IEEE Trans. Inf. Theory, vol. IT-19, no. 4, pp. 434-439, 1973.

[3] E. Parzen, "On the estimation of a probability density function and mode," Ann. Math. Statist., vol. 33, pp. 1065-1076, 1962.

[4] M. D. Richard and R. P. Lippmann, "Neural network classifier estimate Bayesian *a posteriori* probabilities," Neural Computa., vol. 3, pp.461-483, 1991.

[5] A. van Ooyen and B. Nienhuis, "Improving the convergence of the backpropagation algorithm," Neural Networks, vol. 5, pp. 465-471, 1992.

[6] S.-H. Oh, "Improving the error back-propagation algorithm with a modified error function," IEEE Trans. Neural Networks, Vol. 8, pp. 799-803, 1997.

[7] S.-H. Oh, ``Error back-propagation algorithm for classification of imbalanced data," Neurocomputing, Vol. 74, pp. 1058-1061, 2011.