The Korean Communications
in Statistics Vol. 4, No. 3, 1997
pp. 827 - 832

Bootstrap Model Selection Criterion for Determining the Number of Hidden Units in Neural Network Model[1)]

Changha Hwang and Daehak Kim[2)]

Abstract

Statistical relations between a system and empirical distribution are studied in terms of the concept of Akaike's Information Criterion. From this consideration we derive a bootstrap criterion for determining the optimal number of hidden units in neural networks.

1. Introduction

In data analysis and AI research, one of the most important applications of neural networks is modeling a system whose input and output relationship is unknown. In general, weights of the network are modified by means of a stochastic gradient decent method which minimizes a certain error function composed of input and output samples observed from the actual system.

In neural networks, an important but difficult problem is to determine the optimal number of hidden units. This problem is in principle no different from selecting regressors in a linear regression or the order of a polynomial regression. The main idea that have been developed in this area is minimizing some measure of performance over the class of possible models. An increase in the number of the parameters will lessen the output errors for given training samples but will raise additional errors for testing samples. All the candidate measures aim to predict the performance on a set of testing samples, and so to select the model with smallest generalzation error. In this paper, we consider statistical relations between a system and empirical distribution in terms of the concept of Akaike's Information Criterion. From this consideration we derive a bootstrap criterion for determining the optimal number of hidden units in neural networks. Murata *et al.*(1994) proposed network information criterion using the relation between the training error and the generalzation error. Liu(1995) provided a model selection criterion using the cross-validation method to estimate the generalzation error based on the optimal parameter obtained by minimizing an error function. In Statistics there has

1) This research was supported in part by Catholic University of Taegu-Hyosung.
2) Dept. of Statistical Information, Catholic University of Taegu-Hyosung, Kyungbuk, Korea.

been a substantial amount of work in the problem of model selection. Chung et al.(1996) showed recentely that such a bootstrap version of model selection criterion as suggested by Linhart and Zucchini(1986) has a downward bias of amount roughly equivalent to the number of parameters of approximating model.

2. Asymptotic Properties of Learning

Denote the underlying conditional probability distribution as $p(y \mid x)$, the marginal probability distribution of x as $p(x)$, and the joint probability distribution of (x, y) as $p(x, y)$. Hereafter we identify a system with a probability distribution $p(x, y)$. On the other hand, a neural network model such as a feed-forward neural network or a competitive network is denoted by a function $g(x, y, \theta)$, where $\theta \in R^m$ is a parameter vector of weights and thresholds which specify the network. It is modified according to a learning rule. We define a distance function which decides how the model $g(x, y, \theta)$ is closely related to the system $p(x, y)$.

Definition 1. A distance function or the expected loss $D(p, g(\theta))$ between a system p and a model $g(\theta)$ is defined by

$$D(p, g(\theta)) = \int d(x, y, \theta) p(x, y) dx dy$$

where d is a loss plus a regularization term:

$$d(x, y, \theta) = d(p(x, y), g(x, y, \theta)). \quad \blacksquare$$

Our purpose is to find an optimal number of hidden units and the optimal parameter vector, namely the optimal model based on minimizing the distance function given the training data set $\{(x_i, y_i),\ i = 1, \cdots, n\}$. We denote an empirical distribution function as

$$p_n(x, y) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i, y - y_i),$$

where δ is the delta function. It is well known that p_n approximates adequately p if n is large enough. The optimal parameters for the true system and the empirical distribution are denoted by θ_0 and $\hat{\theta}$, respectively:

$$\theta_0 = \arg\min_{\theta} D(p, g(\theta))$$

$$\hat{\theta} = \arg\min_{\theta} D(p_n, g(\theta)).$$

The learning method to be described is obtained by generalizing the back-propagation method or the learning vector quantization method.

Definition 2. The parameter vector θ_t at time t is modified according to the following stochastic gradient descent method:

$$\theta_{t+1} = \theta_t - \varepsilon_t \nabla d(x_t, y_t, \theta_t),$$

where ∇ denotes a differential operator in respect of θ, ε_t is a positive learning coefficient, and t represents the number of modification. Here (x_t, y_t) is an example at time t independently chosen from the training set. ■

In general, the learning coefficient ε_t may be changed depending on time t. But, when the optimal θ_0 deviates slightly, training time would be very long for adjusting the deviation because ε_t might have become too small. In this sense, it is desirable to fix ε_t at a positive small constant. After learning an estimated value $\tilde{\theta}$ is obtained:

$$\tilde{\theta} = \lim_{t\to\infty} \theta_t ,$$

and it varies according to the order of samples applied at the learning period. The probability distribution of θ_t and $\tilde{\theta}$ is denoted by $\pi_t(\theta_t)$ and $\tilde{\pi}(\tilde{\theta})$, repectively. The distribution $\pi_t(\theta_t)$ converges to some distribution $\tilde{\pi}(\tilde{\theta})$. Therefore, when t is large enough, the random variable θ_t is subject to the distribution $\tilde{\pi}(\theta_t)$.

Since the learning rule is stochastic, it is not necessarily guaranteed that θ_t converges to the optimal θ_0, even if we consider only a neighborhood of θ_0 in order to avoid the convergence to a local minimum. Amari(1967, 1993) analyzed the dynamic behavior of θ_t in the neighborhood of θ_0 and obtained the convergence speed to θ_0 and the fluctuation of θ_t around θ_0. In this paper, we show the distribution $\tilde{\pi}(\tilde{\theta})$ approaches the normal distribution with mean θ_0. A brief proof is given in Lemma 1.

Lemma 1. The distribution $\tilde{\pi}(\tilde{\theta})$ approaches the normal distribution

$$N(\theta_0, 2\varepsilon L_Q^{-1} G)$$

as $t\to\infty$. Here, L_Q is a linear operator defined as

$$L_Q M \equiv QM + (QM)^T, \quad M \in R^{m\times m},$$

where $G \equiv V_p[\nabla d(x, y, \theta_0)]$ and $Q \equiv E_p[\nabla\nabla d(x, y, \theta_0)]$.

Proof. After lots of algebra we get the following equation corresponding to the equation (10) of Lemma 2 in Murata *et al.*(1994).

$$\varphi_{n+1}(z) = \varphi_n(z) - \varepsilon z^t Q \frac{\partial}{\partial z}\varphi_n(z) - \varepsilon^2 z^t G z\, \varphi_n(z) + O(\varepsilon^2).$$

Following the next steps in the proof of Lemma 2 in Murata *et al.*(1994) gives us the desired result. ■

Murata *et al.*(1994) showed that the distribution $\tilde{\pi}(\tilde{\theta})$ approaches the normal distribution with mean $\hat{\theta}$. We introduce the result in Lemma 2.

Lemma 2. The distribution $\tilde{\pi}(\tilde{\theta})$ approaches the normal distribution

$$N(\hat{\theta},\ \varepsilon L_{\hat{Q}}^{-1}\hat{G})$$

as $t\to\infty$. Here, $\hat{G}\equiv V_{p_n}[\nabla d(x, y, \hat{\theta})]$ and $\hat{Q}\equiv E_{p_n}[\nabla\nabla d(x, y, \hat{\theta})]$. ■

Let $\hat{\pi}(\hat{\theta})$ be the probability distribution of $\hat{\theta}$. It is known in Statistics that the distribution $\hat{\pi}(\hat{\theta})$ has the following property.

Lemma 3. $\sqrt{n}(\hat{\theta}-\theta_0) \to^d N(0, Q^{-1}GQ^{-1})$ as $n\to\infty$. ■

3. Bootstrap Model Selection Criterion

Since we do not assume that p belongs to the approximating family of models, we use nonparametric bootstrap method. But, naive nonparametric bootstrap may cause some trouble when the sample size is relatively small due to frequent redundancy. First, we choose a bootstrap sample from the empirical cumulative distribution function p_n.

$$(x_1, y_1)^*, \cdots, (x_n, y_n)^* \sim^{iid} p_n.$$

Then, we can compute the bootstrap version of $\tilde{\theta}$ as the minimizer of the bootstrapped empirically discrepancy. It can be known from Chung *et al.*(1996) that bootstrap expected overall discrepancy is given by

$$E^*[D(p_n, g(\tilde{\theta}^*))].$$

In many cases where explicit forms are not easily available, it can be approximated by replacing the basic resampling procedure and taking the average of each bootstrap copies. Formally, with B bootstrap repititions, it can be approximated as

$$B^{-1}\sum_{b=1}^{B} D(p_n, g(\tilde{\theta}_b^*)).$$

Proposition 1. The following approximation holds for the naive plug-in bootstrap expected

overall discrepancy:

$$E^*[D(p_n, g(\tilde{\theta}^*))] = D(p_n, g(\tilde{\theta})) + \frac{1}{n} tr GQ^{-1} + O(n^{-\frac{3}{2}}).$$

Proof. Let us consider the expansion of $D(p_n, g(\tilde{\theta}^*))$ at $\tilde{\theta}$

$$D(p_n, g(\tilde{\theta}^*)) = D(p_n, g(\tilde{\theta})) + (\tilde{\theta}^* - \tilde{\theta})^t \{\nabla D(p_n, g(\tilde{\theta}))\} + \frac{1}{2}(\tilde{\theta}^* - \tilde{\theta})^t \{\nabla\nabla D(p_n, g(\tilde{\theta}))\}(\tilde{\theta}^* - \tilde{\theta}) + \text{h.o.t}.$$

The second term on the right hand side of vanishes, since $\tilde{\theta}$ is chosen that way. Therefore, the expected value of the both sides of the above equation, with respect to bootstrap distribution, turns out to be

$$E^*[D(p_n, g(\tilde{\theta}^*))] = D(p_n, g(\tilde{\theta})) + \frac{1}{2} E^*[(\tilde{\theta}^* - \tilde{\theta})^t \{\nabla\nabla D(p_n, g(\tilde{\theta}))\}(\tilde{\theta}^* - \tilde{\theta})].$$

The second term can be rewritten as

$$\frac{1}{2} E^*[(\tilde{\theta}^* - \tilde{\theta})^t \{\nabla\nabla D(p_n, g(\tilde{\theta}))\}(\tilde{\theta}^* - \tilde{\theta})] \approx \frac{1}{2} E[(\tilde{\theta} - \theta_0)^t \{\nabla\nabla D(p_n, g(\tilde{\theta}))\}(\tilde{\theta} - \theta_0)]$$

$$\approx \frac{1}{2} E_{\tilde{\pi}, \hat{\pi}}[(\tilde{\theta} - \theta_0)^t \{\nabla\nabla D(p_n, g(\tilde{\theta}))\}(\tilde{\theta} - \theta_0)].$$

The integrand on the right hand side can be rewritten as

$$\frac{1}{2}(\tilde{\theta} - \theta_0)^t \{\nabla\nabla D(p_n, g(\tilde{\theta}))\}(\tilde{\theta} - \theta_0)$$

$$= \frac{1}{2} tr[\tilde{Q}\{(\tilde{\theta} - \hat{\theta}) + (\hat{\theta} - \theta_0)\}\{(\tilde{\theta} - \hat{\theta}) + (\hat{\theta} - \theta_0)\}^t]$$

$$= \frac{1}{2} tr[\tilde{Q}\{(\tilde{\theta} - \hat{\theta})(\tilde{\theta} - \hat{\theta})^t + (\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)^t + (\tilde{\theta} - \hat{\theta})(\hat{\theta} - \theta_0)^t + (\hat{\theta} - \theta_0)(\tilde{\theta} - \hat{\theta})^t\}].$$

Averaging each term subject to $\tilde{\pi}(\tilde{\theta})$ and $\hat{\pi}(\hat{\theta})$ and using

$$E_{\tilde{\pi}, \hat{\pi}}[(\tilde{\theta} - \hat{\theta})(\tilde{\theta} - \hat{\theta})^t] = \frac{\varepsilon}{2} \hat{Q}^{-1} \hat{G},$$

$$E_{\tilde{\pi}, \hat{\pi}}[(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)^t] = \frac{1}{n} Q^{-1} G Q^{-1},$$

$$E_{\tilde{\pi}, \hat{\pi}}[(\tilde{\theta} - \hat{\theta})(\hat{\theta} - \theta_0)^t] = 0$$

we get the required result. ∎

Therefore, the bootstrap model selection criterion in neural network is given by

$$D(p_n, g(\tilde{\theta})) + \frac{1}{n} tr \hat{G} \hat{Q}^{-1} .$$

We choose the model minimizes this criterion, which turns out to be equivalent to NIC given by Murata *et al.*(1994).

References

[1] Amari, S. (1967). Theory of adaptive pattern classifiers, *IEEE Transactions on Electronic Computers*, EC-16, 299-307.

[2] Amari, S. (1993). Backpropagation and stochastic gradient descent method, *Neurocomputing*, 5, 185-196.

[3] Chung, H., Lee, K. and Koo, J. (1996). A note on bootstrap model selection criterion, *Statistics and Probability Letters*, 26, 35-41.

[4] Linhart, H. and Zucchini, W.(1986). Model Selection, Wiley, New York.

[5] Liu, Y. (1995). Unbiased estimate of generalization error and model selection in neural network, *Neural Networks*, 8, 215-219.

[6] Murata, N., Yoshizawa, S. and Amari, S. (1991). A criterion for determining the number of parameters in an artificial neural network model, *Artificial Neural Networks*, T. Kohonen, *et al.* ed. Amsterdam: The Netherlands: Elsevier, 9-14.

[7] Murata, N., Yoshizawa, S. and Amari, S. (1994). Network information criterion - Determining the number of hidden units for an artificial neural network model, *IEEE Transactions on Neural Networks*, 5, 865-872.