ESANN'2000 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2000, D-Facto public., ISBN 2-930307-00-5, pp. 69-74

A New Information Criterion for the Selection of Subspace Models

Masashi Sugiyama*, Hidemitsu Ogawa

Department of Computer Science, Tokyo Institute of Technology,
2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan.

Abstract. The problem of model selection is considerably important for acquiring higher levels of generalization capability in supervised learning. In this paper, we propose a new criterion for model selection named the subspace information criterion (SIC). Computer simulations show that SIC works well even when the number of training examples is small.

1. Introduction

Supervised learning is obtaining an underlying rule from training examples, and can be regarded as a function approximation problem. In virtually all learning methods, the quality of the learning results depends heavily on the complexity of *models*.

The problem of model selection has been studied from various standpoints: information statistics [1, 3], Bayesian statistics [5, 2], stochastic complexity [4], and structural risk minimization [6]. Many model selection criteria devised so far use asymptotic approximation in their derivation, so they do not work well when the number of training examples is small

In this paper, we propose a new criterion for model selection from the functional analytic viewpoint. We call this criterion the *subspace information criterion* (SIC). SIC estimates the generalization error by utilizing noise characteristics, instead of asymptotic approximation. Our computer simulations show that SIC works well even when the number of training examples is small.

2. Mathematical foundation of model selection

Let us consider the problem of obtaining an approximation to a target function $f(x)$ of L variables from a set of M training examples. Training examples are made up of input signals $x_m \in \mathcal{D} \subset \mathbf{R}^L$ and corresponding output signals $y_m \in \mathbf{C}$:

$$\{(x_m, y_m) \mid y_m = f(x_m) + n_m\}_{m=1}^{M}, \tag{1}$$

*e-mail: sugi@og.cs.titech.ac.jp, http://ogawa-www.cs.titech.ac.jp/~sugi

ESANN'2000 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2000, D-Facto public., ISBN 2-930307-00-5, pp. 69-74

where y_m is degraded by additive noise n_m. Let θ be a set of factors determining learning results. θ includes, for example, the number and type of basis functions, and parameters in learning algorithms. We call θ a *model*. Let $\hat{f}_\theta$ be a learning result obtained with a model θ. Assuming that f and $\hat{f}_\theta$ belong to a Hilbert space H, the problem of model selection is described as follow.

Definition 1 (Model selection) *From given models, find the one minimizing the generalization error defined as*

$$E_n \|\hat{f}_\theta - f\|^2, \tag{2}$$

where E_n and $\|\cdot\|$ denote the ensemble average over the noise and the norm in H, respectively.

3. Subspace information criterion

In this section, we derive a model selection criterion named the *subspace information criterion* (SIC).

Let y, z, and n be M-dimensional vectors whose m-th elements are y_m, $f(x_m)$, and n_m, respectively:

$$y = z + n. \tag{3}$$

Let X_θ be a mapping from y to $\hat{f}_\theta$:

$$\hat{f}_\theta = X_\theta y. \tag{4}$$

X_θ is called a *learning operator*. In the derivation of SIC, we assume the following conditions.

1. The learning operator X_θ is linear.
2. The mean noise is zero.
3. An unbiased learning result $\hat{f}_u$ has been obtained with a linear operator X_u:
$$E_n \hat{f}_u = f, \quad \hat{f}_u = X_u y. \tag{5}$$

Assumption 1 implies that the range of X_θ becomes a subspace of H. It follows from Eqs.(5), (3), and Assumption 2 that

$$E_n \hat{f}_u = E_n X_u y = E_n X_u z + E_n X_u n = X_u z. \tag{6}$$

Hence, Assumption 3 yields

$$X_u z = f. \tag{7}$$

As shown in the following section, Assumption 3 holds if $M \geq \dim(H)$ under general conditions. The unbiased learning result $\hat{f}_u$ is used for estimating the generalization error of $\hat{f}_\theta$.

ESANN'2000 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2000, D-Facto public., ISBN 2-930307-00-5, pp. 69-74

It is well-known that the generalization error of $\hat{f}_\theta$ is decomposed into the *bias* and *variance*:

$$E_n\|\hat{f}_\theta - f\|^2 = \|E_n\hat{f}_\theta - f\|^2 + E_n\|\hat{f}_\theta - E_n\hat{f}_\theta\|^2. \tag{8}$$

It follows from Eqs.(4) and (3) that Eq.(8) yields

$$E_n\|\hat{f}_\theta - f\|^2 = \|X_\theta z - f\|^2 + \mathrm{tr}\left(X_\theta Q X_\theta^*\right), \tag{9}$$

where $\mathrm{tr}\,(\cdot)$ denotes the trace of an operator, Q is the noise covariance matrix, and X_θ^* denotes the adjoint operator of X_θ. Let X_0 be an operator defined as

$$X_0 = X_\theta - X_u. \tag{10}$$

Then, the bias of $\hat{f}_\theta$ can be expressed by using $\hat{f}_u$ as

$$\|X_\theta z - f\|^2 = \|\hat{f}_\theta - \hat{f}_u\|^2 - 2\mathrm{Re}\langle X_\theta z - f, X_0 n\rangle - \|X_0 n\|^2, \tag{11}$$

where 'Re' stands for the real part of a complex number and $\langle\cdot,\cdot\rangle$ denotes the inner product in H. The second and third terms of the right-hand side of Eq.(11) can not be directly calculated since f and n are unknown. Accordingly, we replace them with the averages of them over the noise. Then, the second term vanishes since the mean noise is zero, and the third term yields

$$E_n\|X_0 n\|^2 = \mathrm{tr}\left(X_0 Q X_0^*\right). \tag{12}$$

Note that this approximation gives an unbiased estimate of the bias:

$$E_n\left(\|\hat{f}_\theta - \hat{f}_u\|^2 - \mathrm{tr}\left(X_0 Q X_0^*\right)\right) = \|X_\theta z - f\|^2. \tag{13}$$

To guarantee that the bias is non-negative, we adopt the following term as an approximation of the bias:

$$\|X_\theta z - f\|^2 \approx \left[\|\hat{f}_\theta - \hat{f}_u\|^2 - \mathrm{tr}\left(X_0 Q X_0^*\right)\right]_+ \quad \text{where} \quad [t]_+ = \max(0, t). \tag{14}$$

Substituting Eq.(14) into Eq.(9), we have the following model selection criterion.

Definition 2 (Subspace information criterion) *Among the given models, select the one minimizing the following SIC:*

$$\mathrm{SIC} = \left[\|\hat{f}_\theta - \hat{f}_u\|^2 - \mathrm{tr}\left(X_0 Q X_0^*\right)\right]_+ + \mathrm{tr}\left(X_\theta Q X_\theta^*\right). \tag{15}$$

The model minimizing SIC is called the *minimum SIC model* (MSIC model), and the learning result obtained by the MSIC model is called the *MSIC learning result*. The generalization capability of the MSIC learning result measured by Eq.(2) is expected to be the best.

ESANN'2000 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2000, D-Facto public., ISBN 2-930307-00-5, pp. 69-74

Unlike well-known Akaike's information criterion (AIC) [1] and the Bayesian information criterion (BIC) [5], SIC estimates the generalization error by utilizing the noise characteristics instead of asymptotic approximation. Therefore, SIC is expected to work well even when the number of training examples is small. Indeed, computer simulations performed in the following section support this claim.

AIC-type criteria are said to be effective only in the selection of nested models [3]. In contrast, no restriction is imposed on models in SIC.

4. Computer simulation

In this section, computer simulations are performed to demonstrate the effectiveness of SIC compared with the network information criterion (NIC) [3], a generalized AIC.

Let the learning target function $f(x)$ be

$$\begin{aligned} f(x) &= \sqrt{2}(\sin x + 2\cos x - \sin 2x - 2\cos 2x + \sin 3x \\ &\quad - \cos 3x + 2\sin 4x - \cos 4x + \sin 5x - \cos 5x), \end{aligned} \tag{16}$$

and training examples $\{(x_m, y_m)\}_{m=1}^{M}$ be

$$x_m = -\pi - \frac{\pi}{M} + \frac{2\pi m}{M}, \quad y_m = f(x_m) + n_m, \tag{17}$$

where the noise n_m is subject to the normal distribution with mean 0 and variance 3. Let us consider the following set of models:

$$\{S_N\}_{N=1}^{20}, \tag{18}$$

where S_N is a Hilbert space spanned by $\{1, \sin nx, \cos nx\}_{n=1}^{N}$, and the inner product is defined as

$$\langle f, g \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)\overline{g(x)}dx. \tag{19}$$

We adopt the least mean squares (LMS) learning aimed at minimizing the *training error*:

$$\sum_{m=1}^{M} \left| \hat{f}_\theta(x_m) - y_m \right|^2. \tag{20}$$

Our task is to find the best model minimizing

$$\text{Error} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \hat{f}_\theta(x) - f(x) \right|^2 dx. \tag{21}$$

Let us consider the following model selection methods for $M = 50$ and $M = 200$.

ESANN'2000 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2000, D-Facto public., ISBN 2-930307-00-5, pp. 69-74

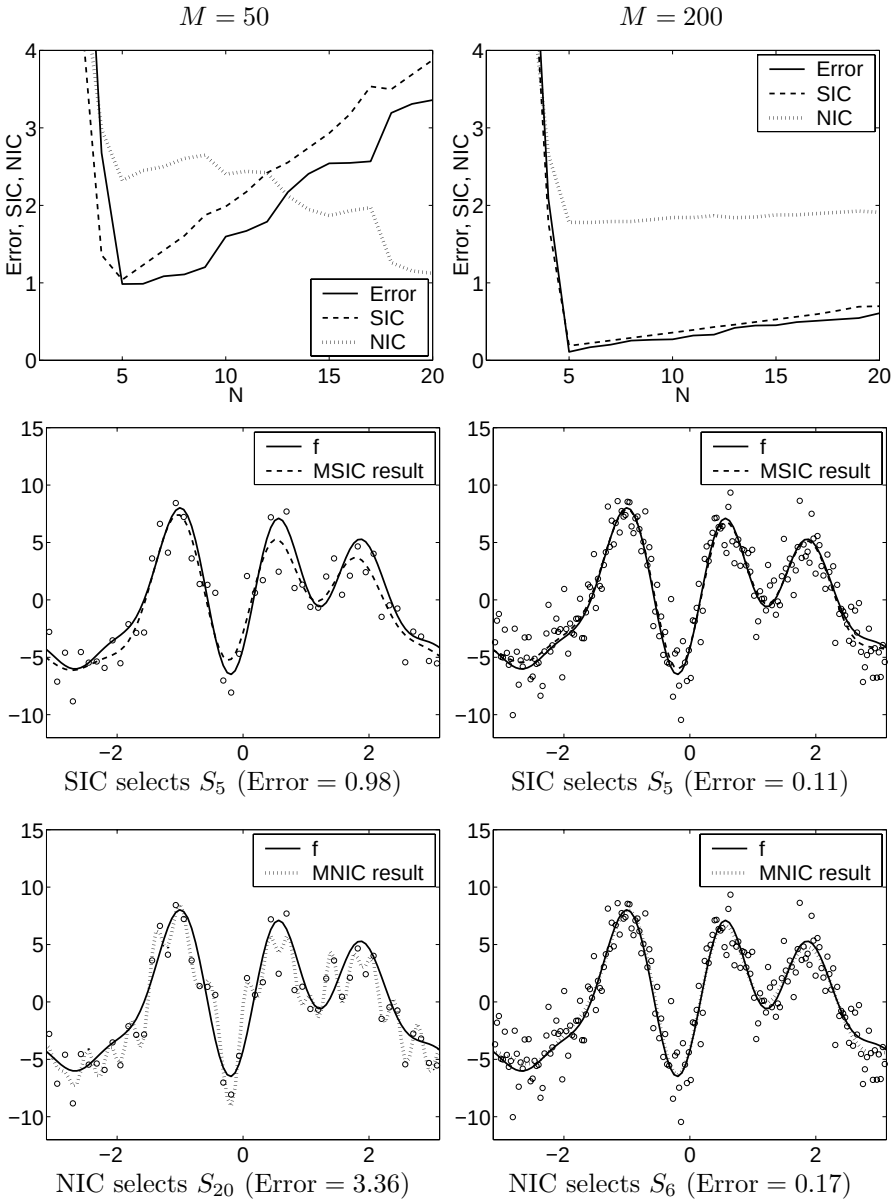

Figure 1: Simulation results when the numbers of training examples are 50 (left) and 200 (right). The top graphs show the values of the error measured by Eq.(21), SIC, and NIC in each model, denoted by the solid, dashed, and dotted lines, respectively. The horizontal axis denotes the highest order N of trigonometric polynomials (see Eq.(18)). The middle and bottom graphs show the target function $f(x)$ (solid line), training examples ('$\circ$'), MSIC and minimum NIC (MNIC) learning results (dashed lines).

ESANN'2000 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2000, D-Facto public., ISBN 2-930307-00-5, pp. 69-74

(A) SIC: Let $H = S_{20}$ which includes all models. Since the learning result obtained with S_{20} is unbiased, it is adopted as $\hat{f}_u$. The noise covariance matrix Q is estimated by assuming $Q = \sigma^2 I$ and estimating σ^2 as

$$\hat{\sigma}^2 = \sum_{m=1}^{M} \left| \hat{f}_u(x_m) - y_m \right|^2 \Big/ (M - \dim(H)) \,. \tag{22}$$

(B) NIC: The squared loss is adopted as the loss function. The distribution of sample points given by Eq.(17) is regarded as a uniform distribution.

In both (A) and (B), no a priori information is used and the LMS estimator is commonly adopted. Hence, the efficiency in these model selection methods can be fairly compared by this simulation.

Fig.1 shows the simulation results. These results show that when $M = 200$, both SIC and NIC give reasonable learning results. However, when it comes to the case when $M = 50$, SIC outperforms NIC. This implies that SIC works well even when the number of training examples is small.

5. Conclusion

We proposed a new model selection criterion called the subspace information criterion (SIC). In SIC, the generalization error is estimated by utilizing the noise characteristics instead of asymptotic approximation. Computer simulations showed that SIC works well even when the number of training examples is small.

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19(6):716–723, 1974.

[2] D. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

[3] N. Murata, S. Yoshizawa, and S. Amari. Network information criterion—determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks*, 5(6):865–872, 1994.

[4] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[5] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[6] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin, 1995.