VC dimension theory for a learning system with forgetting

Shotaro Akaho

Mathematical Informatics Sc.
Electrotechnical Laboratory
1-1-4, Umezono, Tsukuba-shi, Ibaraki 305, Japan

Abstract In a changing environment, forgetting old samples is an effective method to improve the adaptability of learning systems. However, too fast forgetting causes a decrease of generalization performance. In this paper, we analyze the generalization performance of a learning system with a forgetting parameter. For a class of binary discriminant functions, it is proved that the generalization error is given by $O(\sqrt{h\alpha})$ ($O(h\alpha)$ in a certain case), where h is the VC dimension of the class of functions and $1-\alpha$ represents a forgetting rate. The result provides a criterion to determine the optimal forgetting rate.

1 Introduction

Generalization performance of neural networks is one of the most important issue, and it has been a subject of interest to many researchers. Baum and Haussler presented the generalization performance of neural networks using Vapnik-Chervonenkis(VC) dimension[1].

A neural network adapts itself to samples given from outer environment. However, even if the network would achieve a small risk for the training samples, it might not always fit unknown test samples well (*generalization problem*).

We call generalization error the difference between the actual risk for the whole set of data and the empirical risk for training samples. Roughly speaking, the generalization error reduces, as the number of samples increases and the capacity of the network decreases.

When a lot of samples are given, however, there are some cases in which we had not better use the whole samples. For instance:

(1) when the environment changes gradually, old samples are not so reliable as newer ones,

(2) when the capacity of network is fixed, a lot of samples may leads to capacity overflow.

In such cases, the empirical risk will become large, though the generalization error is small. Hence some mechanism to forget old samples is needed.

In this paper, we consider the learning system that minimizes the risk in which old samples are weighted exponentially small, and analyze its generalization performance.

2 Generalization problem and VC dimension

Let us formulate the generalization problem as expected risk minimization problem[2]. For the sake of simplicty, we consider only the learning of binary discriminant function.

A sample is a pair of an input vector $\boldsymbol{x}$ and a class $y \in \{0,1\}$ to which $\boldsymbol{x}$ belongs, and training samples $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)$ are randomly and independently generated subject to a certain unknown probability distribution $P(\boldsymbol{x}, y)$.

Let us consider a class of neural networks with parameter $\boldsymbol{w}$. A neural network calculates an indicator function $f(\boldsymbol{x}, \boldsymbol{w}) \in \{0,1\}$ for input $\boldsymbol{x}$.

The ultimate goal of the learning is minimizing the expected risk $R(\boldsymbol{w})$, which is the probability of misclassification in this case,

$$R(\boldsymbol{w}) = \int (y - f(\boldsymbol{x}, \boldsymbol{w}))^2 P(\boldsymbol{x}, y)\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}y \tag{1}$$

by using only training samples. Let $\boldsymbol{w}_0$ be the parameter that minimizes $R(\boldsymbol{w})$. Since $P(\boldsymbol{x}, y)$ is unknown, we cannot get the exact value of $\boldsymbol{w}_0$. Thus usually the empirical risk $R^*_{\mathrm{emp}}(\boldsymbol{w})$, which is the frequency of errors for the training samples in this case,

$$R^*_{\mathrm{emp}}(\boldsymbol{w}) = \frac{1}{m} \sum_{i=1}^{m} (y_i - f(\boldsymbol{x}_i, \boldsymbol{w}))^2 \tag{2}$$

is minimized instead of $R(\boldsymbol{w})$. Let $\boldsymbol{w}^*_{\mathrm{emp}}$ be the parameter that minimizes $R^*_{\mathrm{emp}}(\boldsymbol{w})$.

Generalization error is the difference between the expected risk of $\boldsymbol{w}_0$ and that of $\boldsymbol{w}^*_{\mathrm{emp}}$,

$$D = R(\boldsymbol{w}^*_{\mathrm{emp}}) - R(\boldsymbol{w}_0). \tag{3}$$

Vapnik et al. analyzed the distribution of D. In his theory, the VC dimension, which represents the complexity or the capacity of a network, plays an essential role.

Definition 1 The VC dimension *of a set of binary functions* $\{f(\boldsymbol{x}, \boldsymbol{w})\}$ *is defined by the maximal number* h *of vectors* $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_h\}$ *which can be dichotomized in all possible* 2^h *ways by using functions in the set.*

As a model of forgetting old samples, we consider "the (exponentially) weighted empirical risk" $R_{\mathrm{emp}}(\boldsymbol{w})$ instead of $R^*_{\mathrm{emp}}(\boldsymbol{w})$. For the sake of simplicity, let the number of samples be infinite.

Definition 2 *Let* t *be the current time, and assume that each sample* $(\boldsymbol{x}_i, y_i)$ *is given at time* $t - i$. **The weighted empirical risk** *is defined by*

$$R_{\mathrm{emp}}(\boldsymbol{w}) = \frac{1}{A} \sum_{i=0}^{\infty} (1 - \alpha)^i (y_i - f(\boldsymbol{x}_i, \boldsymbol{w}))^2, \tag{4}$$

where $1 - \alpha$ *represents a forgetting rate and* A *is a normalization parameter* ($A = 1/\alpha$).

3 Main result

In order to analyze the distribution of generalization error, we use the uniform convergence technique established by Vapnik[3].

Assume that the inequality

$$\Pr\left[\sup_{\boldsymbol{w}} |R_{\mathrm{emp}}(\boldsymbol{w}) - R(\boldsymbol{w})| > \kappa\right] < \delta \tag{5}$$

holds, it follows

$$\Pr\left[R(\boldsymbol{w}_{\mathrm{emp}}) - R(\boldsymbol{w}_0) > 2\kappa\right] < \delta, \tag{6}$$

where $\boldsymbol{w}_{\mathrm{emp}}$ and $\boldsymbol{w}_0$ are parameters that minimize $R_{\mathrm{emp}}(\boldsymbol{w})$ and $R(\boldsymbol{w})$ respectively. This inequality represents the confidence interval of the generalization error.

The following two theorems give upper bounds of κ for a given δ.

Theorem 1 *Let $\{f(\boldsymbol{x}, \boldsymbol{w})\}$ be a class of binary functions with VC dimension h, and let $R_{\rm emp}(\boldsymbol{w})$ be the weighted empirical risk for samples that is defined in equation (4). Then for a sufficientlly small α, the actual risk is bounded with probability $1-\delta$ by*

$$R(\boldsymbol{w}) < R_{\rm emp}(\boldsymbol{w}) + \sqrt{\frac{h\log 2e - \log\delta}{4(2-\alpha)}\alpha} + 4\sqrt{2}\alpha(1-\alpha)^h. \tag{7}$$

When the empirical risk $R_{\rm emp}(\boldsymbol{w})$ is small, the following may give a better bound.

Theorem 2 *Let $\{f(\boldsymbol{x}, \boldsymbol{w})\}$ be a class of binary functions with VC dimension h, and let $R_{\rm emp}(\boldsymbol{w})$ be the weighted empirical risk for samples that is defined in equation (4). Then for a sufficientlly small α, the actual risk is bounded with probability $1-\delta$ by*

$$R(\boldsymbol{w}) < R_{\rm emp}(\boldsymbol{w}) + \frac{h\log 2e - \log\delta}{2(2-\alpha)}\alpha\left(1 + \sqrt{1 + \frac{4(2-\alpha)}{h\log 2e - \log\delta}R_{\rm emp}(\boldsymbol{w})}\right) + 4\sqrt{2}\alpha(1-\alpha)^h. \tag{8}$$

Theorem 1 and 2 give us criteria to determine forgetting parameter α. From the theorems, we can estimate the confidence interval of the minimal value of the expected risk by

$$R(\boldsymbol{w}_0) < R_{\rm emp}(\boldsymbol{w}_{\rm emp}) + C(h, \alpha, \delta), \tag{9}$$

where $C(h, \alpha, \delta)$ is decided by (7) or (8). We shall take such α that minimizes this bound.

4 Mathematical Analysis

In this section, we show the outline of proofs of the theorems described in the section above.

We shall estimate the following value,

$$X = \Pr\left[\sup_{\boldsymbol{w}} |R_{\rm emp}(\boldsymbol{w}) - R(\boldsymbol{w})| > \kappa\right]. \tag{10}$$

From the central limit theorem, the distribution of $R_{\rm emp}(\boldsymbol{w})$ can be approximated by the normal distribution with mean $R(\boldsymbol{w})$ and variance $R(\boldsymbol{w})(1-R(\boldsymbol{w}))$ when α is sufficiently small.

By this approximation, we get

$$X \simeq \Pr\left[\sup_{\boldsymbol{w}} |R_{\rm emp1}(\boldsymbol{w}) - R_{\rm emp2}(\boldsymbol{w})| > \frac{\kappa}{\sqrt{2}}\right], \tag{11}$$

where $R_{\rm emp1}(\boldsymbol{w})$ and $R_{\rm emp2}(\boldsymbol{w})$ are the values of $R_{\rm emp}(\boldsymbol{w})$ for two independent experiments.

In order to evaluate the value above, let us introduce the partial sum of weighted risk,

$$R^{(l)}_{\rm emp}(\boldsymbol{w}) = \frac{1}{A}\sum_{i=0}^{l}(1-\alpha)^i(y_i - f(\boldsymbol{x}_i, \boldsymbol{w}))^2. \tag{12}$$

Using this value and the fact that $\sup_i z_i < \sum_i z_i$ if $z_i \geq 0$ for all i, we have

$$X < \sum_{\boldsymbol{w}\in W} \Pr\left[\left|R^{(l)}_{\rm emp1}(\boldsymbol{w}) - R^{(l)}_{\rm emp2}(\boldsymbol{w})\right| > \frac{\kappa}{\sqrt{2}} - 2\alpha(1-\alpha)^l\right], \tag{13}$$

where W is a set of $\boldsymbol{w}$ for all possible values of $(R^{(l)}_{\rm emp1}(\boldsymbol{w}), R^{(l)}_{\rm emp2}(\boldsymbol{w}))$, and it follows

$$X < \sum_{\boldsymbol{w}\in W} \Pr\left[|R_{\rm emp}(\boldsymbol{w}) - R(\boldsymbol{w})| > \kappa'\right], \tag{14}$$

where $\kappa' = \kappa - 4\sqrt{2}\alpha(1-\alpha)^l$.

Vapnik et al.[3] has shown that the number of all possible values of $((y_1 - f(\boldsymbol{x}_1, \boldsymbol{w}))^2, \ldots, (y_l - f(\boldsymbol{x}_l, \boldsymbol{w}))^2)$ is bounded by $1.5l^h/h!$ for $l > h$, where h is the VC dimension. Thus we get

$$\text{The number of elements in } W < 1.5\frac{(2l)^h}{h!}. \tag{15}$$

For each fixed $\boldsymbol{w}$, we can show the following inequality,

$$\Pr\left[|R_{\text{emp}}(\boldsymbol{w}) - R(\boldsymbol{w})| > \kappa\right] < \exp\left(-\frac{4(2-\alpha)}{\alpha}\kappa^2\right). \tag{16}$$

Applying (16) and (15) to (14), we obtain

$$X < 1.5\frac{(2l)^h}{h!}\exp\left(-\frac{4(2-\alpha)}{\alpha}\kappa'^2\right). \tag{17}$$

Solving the equation that the right hand side of the inequality (17) is equal to δ, and putting $l = h$, we arrive at the result of theorem 1.

Similarly, by showing the following inequality, we conclude theorem 2.

$$\Pr\left[\frac{|R_{\text{emp}}(\boldsymbol{w}) - R(\boldsymbol{w})|}{\sqrt{R(\boldsymbol{w})}} > \kappa\right] < \exp\left(-\frac{2-\alpha}{\alpha}\kappa^2\right). \tag{18}$$

5 Concluding remarks and future works

In this paper, we have established the generalization theory for the learning system that minimizes the weighted risk based on VC dimension. Theorem 1 and 2 give us criteria to determine the forgetting parameter, but we must still consider the fast algorithm of the determination.

VC dimension can be defined not only for a class of binary functions, but also for all parameterized functions, such as continuous functions and unsupervised learning (no teacher signal for output). Our result can be generalized for those learning.

Many criteria have been proposed for network selection such as AIC and MDL. We will also be able to reformulate those criteria in terms of forgetting.

In our result, the change of environment is not explicitly discussed, i.e., samples are assumed to be given from the same distribution and the change of environment causes just an increase of the empirical risk. However, its explicit treatment will be very difficult.

References

[1] E.B. Baum and D. Haussler: What size net gives valid generalization? *Neural Computation*, Vol. 1, pp. 151–160, 1989.

[2] V.A. Vapnik: *Estimation of Dependences Based on Empirical Data.* Springer-Verlag, 1984.

[3] V.N. Vapnik and A.Ya. Chervonenkis: On the uniform convergence of relative frequencies of events to their probabilities. *Theory Prob. Appl.*, Vol. 16, pp. 264–280, 1971.