GENERALIZATION BOUNDS FOR THE REGRESSION OF REAL-VALUED FUNCTIONS

Rhee Man Kil and Imhoi Koo
Division of Applied Mathematics
Korea Advenced Institute of Science and Technology

Abstract

This paper suggests a new bound of estimating the confidence interval defined by the absolute value of difference between the true (or general) and empirical risks for the regression of real-valued functions using incremental learning algorithms. The theoretical bounds of confidence intervals can be derived in the sense of probably approximately correct (PAC) learning. However, these theoretical bounds are too overestimated and not well fitted to the empirical data. In this sense, a new bound of the confidence interval which can explain the behavior of learning machines using incremental learning more faithfully to the given samples, is suggested.

Keywords : VC dimension, confidence interval, generalization, regression, incremental learning

1 INTRODUCTION

This paper suggests a new bound of the confidence interval defined by the absolute value of difference between the true (or general) and empirical risks for the regression of real-valued functions. In the learning models, the goal is minimizing the general error for the whole distribution of sample space, not just a set of training samples. This is refer to as the generalization problem. To identify the general error, we need to estimate the confidence interval. Usually, the validation set, a part of training samples is extracted and used to estimate the confidence interval. But the physical model of confidence intervals is not well known.

The regression of real-valued functions is frequently tackled by the incremental learning algorithms in which the necessary computational units, usually the kernel functions with locality[1, 2, 3, 4, 5] such as Gaussian kernel functions, are recruited in the learning procedure. In the computational learning theory, the bounds of confidence intervals for regression can be derived in the sense of probably approximately correct (PAC) learning[6]. For a network with Gaussian kernel functions, the upper bounds of confidence intervals can be derived using the concept of covering numbers[7] of the hypothesis space defined by network models. The covering number, in a sense, indicates the mapping capability of network models. The final form of confidence intervals[8, 9] is a function of the numbers of input dimension, kernel functions, and training samples. However, these theoretical bounds are too overestimated and not well fitted to the empirical data. In this sense, a new bound of the confidence interval which can explain the behavior of general error more faithfully to the given samples, is suggested. The suggested bound of the confidence interval can be applied to the optimization of learning models in the sense of minimizing the general error.

2 THE CONFIDENCE INTERVALS OF RISK FUNCTIONS FOR REGRESSION

Let us consider that the following l samples

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_l, y_l)$$

are formed randomly and independently. And we assume that $\mathbf{x} \in X$ (input space) and $y \in Y$ (output space) have the functional relationship $y = f(\mathbf{x})$.

Now let us consider the vector $\mathbf{x}$ appear randomly and independently in accordance with a distribution $F(\mathbf{x})$. Then the values of y are generated from the random trials in accordance with $F(y|\mathbf{x})$. Here, we define a loss function as

$$Q(\mathbf{x}, n) = (f(\mathbf{x}) - f_n(\mathbf{x}))^2 \tag{1}$$

where $f(\mathbf{x})$ and $f_n(\mathbf{x})$ represent the target function and the estimation function in F_n, the hypothesis space (or structure) with n parameters respectively.

The goal of learning models (or estimation function) f_n is minimizing the true (or general) risk defined by

$$R(f_n) = E[Q(\mathbf{x}, n)] = \int (f(\mathbf{x}) - f_n(\mathbf{x}))^2 dF(y|\mathbf{x}). \tag{2}$$

However, we can not estimate (2) in advance. Instead, we get the empirical risk R_{emp} for l samples defined by

$$R_{emp}(f_n) = \frac{1}{l}\sum_{i=1}^{l}(f(\mathbf{x}_i) - f_n(\mathbf{x}_i))^2 \tag{3}$$

where $\mathbf{x}_i$ and $f(\mathbf{x}_i)$ represent the ith input and output samples respectively.

Let us also define a random variable

$$e_n(\mathbf{x}) = f(\mathbf{x}) - f_n(\mathbf{x}). \tag{4}$$

Then e_n represents a random variable associated with the error for the estimator f_n. For the case that e_n has unknown distribution, Vapnik[10] derives the bounds of the confidence intervals of risk functions for the problems of classification and regression. Since these bounds accounts for general cases, in other words, unknown distribution of e_n, unknown estimation function f_n, and unknown learning algorithm, they are usually too overestimated in a specific case of learning model. In this sense, a new bound of the confidence interval which can explain the behavior of learning machines more faithfully to the given samples, is investigated.

First, let us analyze the relationship between the true and empirical risks in the sense of the distribution of e_n. Especially, we consider the confidence interval defined by the absolute value of difference between the true and empirical risks. That is, the confidence interval $R_c(f_n)$ is defined by

$$R_c(f_n) = |R(f_n) - R_{emp}(f_n)|. \tag{5}$$

For the confidence interval of (5), we propose the following theorem.

Theorem 1 *The confidence intervals of risk functions are bounded by the following inequality with the probability at least $1-\delta$:*

$$|R(f_n) - R_{emp}(f_n)| < \sqrt{\frac{k-1}{l\delta}} R(f_n) \tag{6}$$

where k is a constant dependent upon the distribution of estimation errors. In the case of normal distribution $k = 3$, otherwise, k is a constant greater than or equal to 1.

(proof)

Let us assume that the expected value of $e_n(\mathbf{x})$ of (4) is given by

$$E[e_n(\mathbf{x})] = \eta. \tag{7}$$

Then the variance of $y_n(\mathbf{x})$ is given by

$$Var[e_n(\mathbf{x})] = E[e_n^2(\mathbf{x})] - E^2[e_n(\mathbf{x})] = R(f_n) - \eta^2. \tag{8}$$

Let us define another random variable

$$z_n(\mathbf{x}) = (f(\mathbf{x}) - f_n(\mathbf{x}))^2 = e_n^2(\mathbf{x}). \tag{9}$$

Then the expected value of $z_n(\mathbf{x})$ is given by

$$E[z_n(\mathbf{x})] = E[e_n^2(\mathbf{x})] = R(f_n) \tag{10}$$

and the variance of $z_n(\mathbf{x})$ is given by

$$Var[z_n(\mathbf{x})] = E[z_n^2(\mathbf{x})] - E^2[z_n(\mathbf{x})]. \tag{11}$$

Here, we assume that there exists a constant k such that

$$\frac{E[z_n^2(\mathbf{x})]}{E^2[z_n(\mathbf{x})]} \leq k, \tag{12}$$

then

$$Var[z_n(\mathbf{x})] \leq (k-1)E^2[z_n(\mathbf{x})] = (k-1)R^2(f_n). \tag{13}$$

In general, the constant k is a constant greater than or equal to 1 since the variance of $z_n(\mathbf{x})$ is always nonnegative.

If the random variable $e_n(\mathbf{x})$ in (4) has Gaussian probability density function,

$$\begin{aligned} E[z_n^2(\mathbf{x})] &= \frac{1}{\sqrt{2\pi}\sigma_{e_n}} \int_{-\infty}^{+\infty} e_n^4(\mathbf{x}) e^{-\frac{(e_n(\mathbf{x})-\eta)^2}{2\sigma_{e_n}^2}} de_n(\mathbf{x}) \\ &= (3\sigma_{e_n}^2 + \eta^2)(\sigma_{en}^2 + \eta^2) \\ &\leq 3R^2(f_n). \end{aligned} \tag{14}$$

In other words,

$$Var[z_n(\mathbf{x})] \leq 2R^2(f_n). \tag{15}$$

Let us also define a random variable $\bar{z}_n(\mathbf{x})$ which is an average of l $z_n(\mathbf{x})$s',

$$\bar{z}_n = \frac{1}{l}\sum_{i}^{l} z_n(\mathbf{x}_i) = R_{emp}(f_n). \tag{16}$$

Then the expected value of $\bar{z}_n$ is given by

$$E[\bar{z}_n] = E[z_n(\mathbf{x})] = R(f_n) \tag{17}$$

and the variance of $\bar{z}_n$ is given by

$$Var[\bar{z}_n] = \frac{1}{l} Var[z_n(\mathbf{x})] = \frac{k-1}{l} R^2(f_n) \tag{18}$$

assuming that $z_n(\mathbf{x}_i)$s'are uncorrelated each other.

The random variable $\bar{z}_n$ satisfies the following Tchebycheff inequality[11]:

$$\Pr[|E[\bar{z}_n] - \bar{z}_n(\mathbf{x})| < \epsilon_n] = \Pr[|R(f_n) - R_{emp}(f_n)| < \epsilon_n] \geq 1 - \frac{Var[\bar{z}_n]}{\epsilon_n^2} \tag{19}$$

regardless of the distribution of e_n.

For the PAC learning, we set

$$\delta = \frac{Var[\bar{z}_n]}{\epsilon_n^2} = \frac{k-1}{\epsilon_n^2 l} R^2(f_n). \tag{20}$$

Then the bound of ϵ_n becomes

$$\epsilon_n = \sqrt{\frac{k-1}{l\delta}} R(f_n). \tag{21}$$

Therefore, with the probability at least $1-\delta$,

$$|R(f_n) - R_{emp}(f_n)| < \sqrt{\frac{k-1}{l\delta}} R(f_n). \tag{22}$$

Q.E.D.

In general, the inequality

$$R_{emp}(f_n) \leq R(f_n) \tag{23}$$

holds in the sense of empirical risk minimization. Therefore, we can easily conclude the following corollary from theorem 1:

Corollary 1 *The true risks have the following upper bounds with the probability at least* $1-\delta$*:*

$$R(f_n) < (1 - \sqrt{\frac{k-1}{l\delta}})^{-1} R_{emp}(f_n). \tag{24}$$

For the nonnegative loss functions with the finite VC dimension, Vapnik suggests the following bounds on the generalization for regression[10]:

Theorem 2 (Vapnik, 1998) *For the nonnegative loss functions with the finite VC dimension* h_n*, the risk functions of* f_n *are bounded by the following inequality with the probability at least* $1-\delta$*:*

$$R(f_n) \leq (1 - \tau_p a(p)\sqrt{\varepsilon_n})_+^{-1} R_{emp}(f_n) \ \ \textit{for } p > 2, \tag{25}$$

$$\sup_{f_n \in F_n} \frac{(E[z_n^p])^{1/p}}{E[z_n]} = \tau_p, \tag{26}$$

$$a(p) = (\frac{1}{2}(\frac{p-1}{p-2})^{p-1})^{1/p}, \ \ \textit{and} \tag{27}$$

$$\varepsilon_n = 4\frac{h_n(1 + \ln\frac{2l}{h_n}) - \ln\frac{\delta}{4}}{l} \tag{28}$$

for any $f_n \in F_n$*.*

Comparing the theorems 1 and 2, we can see that the true risks are bounded by a multiplication of some constant and the empirical risk. The multiplication terms in the theorems 1 and 2 are mainly dependent upon $\sqrt{1/l}$ and $\sqrt{h_n/l}$ respectively. In other words, the bounds of true risks in the theorem 1 are determined by the multiplication term mainly dependent upon the number of samples while the bounds of true risks in the theorem 2 are determined by the multiplication term mainly dependent upon the ratio of the VC dimension and the number of samples.

Here, we need to investigate the property of true risks to obtain the bounds of confidence intervals. First, let us consider the case of the bounded loss functions with the finite VC dimension, in other words,

$$0 \leq Q(\mathbf{x}, n) \leq B. \tag{29}$$

For this case, Vapnik suggests the following bounds on the generalization for regression[10]:

Theorem 3 (Vapnik, 1998) *For the bounded loss functions with the finite VC dimension h_n, the risk functions of f_n are bounded by the following inequality with the probability at least $1-\delta$:*

$$R(f_n) \leq R_{emp}(f_n) + \frac{B\varepsilon_n}{2}(1 + \sqrt{1 + \frac{4R_{emp}(f_n)}{B\varepsilon_n}}) \text{ and} \tag{30}$$

$$\varepsilon_n = 4\frac{h_n(1+\ln\frac{2l}{h_n}) - \ln\frac{\delta}{4}}{l} \tag{31}$$

for any $f_n \in F_n$.

The generalization bounds of the true risks for the bounded loss functions can be explained by the following theorem:

Theorem 4 *For the bounded loss functions with the finite VC dimension h_n, the true risks have the following upper bounds with the probability at least $1-2\delta$:*

$$R(f_n) \leq C_1 r_n + C_2 \epsilon_n, \tag{32}$$

$$r_n = (\frac{1}{n})^{\alpha}, \tag{33}$$

$$\epsilon_n = \frac{B\varepsilon_n}{2}(1 + \sqrt{1 + \frac{4R_{emp}(f_n)}{B\varepsilon_n}}), \text{ and} \tag{34}$$

$$\varepsilon_n = 4\frac{h_n(1+\ln\frac{2l}{h_n}) - \ln\frac{\delta}{4}}{l} \tag{35}$$

where k is a constant dependent upon the distribution of estimation errors, C_1 is a constant dependent upon the target function, C_2 is a constant dependent upon estimation error, and α is a constant dependent upon the smoothness of target function and the dimension of input space.

(proof)

The true risk in (6) can be rewritten as

$$R(f_n) = R(f_n) - R(f_n^*) + R(f_n^*) \tag{36}$$

where $R(f_n^*)$ is the optimal estimation function in F_n.

Then using the triangular inequality,

$$R(f_n) \leq |R(f_n) - R(f_n^*)| + R(f_n^*). \tag{37}$$

Let us first consider the first term $|R(f_n) - R(f_n^*)|$ of the inequality (37). In the case of empirical risk minimization, the following conditions holds:

$$\begin{aligned} R(f_n^*) &\leq R(f_n) \qquad (38) \\ R_{emp}(f_n^*) &\geq R_{emp}(f_n). \qquad (39) \end{aligned}$$

Then according to the theorem 3, with the probability at least $1 - 2\delta$,

$$|R(f_n) - R_{emp}(f_n)| \leq \epsilon_n \ \text{ for any } f_n \in F_n \tag{40}$$

where

$$\epsilon_n = \frac{B\varepsilon_n}{2}(1 + \sqrt{1 + \frac{4R_{emp}(f_n)}{B\varepsilon_n}}). \tag{41}$$

Therefore, with the probability at least $1 - 2\delta$,

$$|R(f_n) - R(f_n^*)| \leq 2\epsilon_n \tag{42}$$

to satisfy the conditions of (38) and (39).

The second term $R(f_n^*)$ of the inequality (37) is concerned with the theorem of function approximation. According to Lorentz's work[12], if n represents the degree of polynomials, the true risk of f_n^* is bounded by

$$R(f_n^*) \leq r_n = N(\frac{1}{n})^{s/m} \tag{43}$$

where s represents the number of existing derivatives, m represents the dimension of input space of the target function, and N represents a constant that depends on the target function and s.

In the case that the estimation function is composed of sigmoid or Gaussian kernel functions, r_n is bounded[13, 14, 15] by

$$r_n \leq O(\frac{1}{\sqrt{n}}). \tag{44}$$

Therefore, from (37), (42) and (43), with the probability at least $1 - 2\delta$,

$$R(f_n) \leq C_1(\frac{1}{n})^{\alpha} + C_2\epsilon_n \tag{45}$$

where

$$C_1 = N, \ \ C_2 = 2, \ \text{ and } \ \alpha = s/m. \tag{46}$$

Q.E.D.

Theorem 4 states that the bounds of true risks are determined by the approximation error r_n of (32) and the regression error ϵ_n of (32). Note that the regression error ϵ_n can be approximated as $B\varepsilon_n$ if $R_{emp}(f_n) \ll B\varepsilon_n/4$. For the bounded loss functions with the finite VC dimension, the confidence intervals of risk functions can be described by the following theorem:

Theorem 5 *For the bounded loss functions with the finite VC dimension h_n, the confidence intervals of risk functions have the following upper bounds with the probability at least $1 - 3\delta$:*

$$|R(f_n) - R_{emp}(f_n)| < \sqrt{\frac{k-1}{l\delta}}(C_1 r_n + C_2\epsilon_n), \tag{47}$$

$$r_n = (\frac{1}{n})^{\alpha}, \tag{48}$$

$$\epsilon_n = \frac{B\varepsilon_n}{2}(1 + \sqrt{1 + \frac{4R_{emp}(f_n)}{B\varepsilon_n}}), \ \ \textit{and} \tag{49}$$

$$\varepsilon_n = 4\frac{h_n(1 + \ln\frac{2l}{h_n}) - \ln\frac{\delta}{4}}{l} \tag{50}$$

where k is a constant dependent upon the distribution of estimation errors, C_1 is a constant dependent upon the target function, C_2 is a constant dependent upon estimation error, and α is a constant dependent upon the smoothness of target function and the dimension of input space.

(proof)

The confidence intervals of risk functions are bounded by the following inequality with the probability at least $1-\delta$:

$$|R(f_n) - R_{emp}(f_n)| < \sqrt{\frac{k-1}{l\delta}} R(f_n) \tag{51}$$

from the theorem 1. The true risks have the following upper bound with the probability at least $1-2\delta$:

$$R(f_n) \leq C_1 r_n + C_2 \epsilon_n \tag{52}$$

from the theorem 4. Therefore, from (51) and (52), the confidence intervals of risk functions have the upper bounds with the probability at least $1-3\delta$:

$$|R(f_n) - R_{emp}(f_n)| < \sqrt{\frac{k-1}{l\delta}} (C_1 r_n + C_2 \epsilon_n). \tag{53}$$

Q.E.D.

Theorem 5 states that the confidence intervals of risk functions for the regression of the bounded target functions and the learning models with the finite VC dimension, are bounded by the approximation error r_n of (32) and the regression error ϵ_n of (32), but these bounds are decreased at the rate of $1/\sqrt{l}$ as the number of samples increases. Therefore, the true and empirical risks approach to the same value regardless of the ratio h_n/l as the number of samples increases. This result can be compared with the Vapnik's result of generalization bounds for regression[10] in which the bounds of confidence intervals are proportional to ϵ_n, in other words, approximately h_n/l. In fact, it is commonly observed that the confidence intervals are decreasing when the number of kernel functions is proportional to l and l is decreasing.

There are many algorithms of incremental learning[1, 2, 3, 4, 5] using the kernel functions with locality. In most cases, a learning algorithm composed of two learning processes, is considered. They are the process of recruiting the necessary number of kernel functions and the process of parameter estimation associated with kernel functions. For the estimation of true risks of incremental learning, we can consider the estimation model motivated from the theorem 5. We have shown that such estimation model of confidence intervals[16] could be well fitted to the empirical data.

3 CONCLUSION

In this paper, the confidence intervals for the regression of real-valued functions using incremental learning, are suggested in the sense of VC dimension and function approximation theories. The suggested confidence intervals can be applied to various types of learning models including artificial neural networks for optimizing the structure (or network size) of learning models in the sense of minimizing the general error.

Acknowledgement: This research was supported by *Brain Science and Engineering Research Program* supported by Korean Ministry of Science and Technology.

References

[1] M. Niranjan and F. Fallside. Neural networks and radial basis functions in classifying static speech patterns. Technical Report CUED/F-INFENG/TR22, Cambridge University, 1988.

[2] J. Moody and C. J. Darken. Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1:281–294, 1989.

[3] E. J. Hartman, J. D. Keeler, and J. M. Kowalski. Layered neural networks with gaussian hidden units as universal approximations. *Neural Computation*, 2:210–215, 1990.

[4] S. Lee and R. M. Kil. A gaussian potential function network with hierarchically self-organizing learning. *Neural Networks*, 4(2):207–224, 1991.

[5] R. M. Kil. Function approximation based on a network with kernel functions of bounds and locality: An approach of non-parametric estimation. *ETRI journal*, 15(2):35–51, 1993.

[6] L. Valiant. A theory of learnability. *Communications of ACM*, 27(11):1134–1142, 1984.

[7] D. Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. Technical Report UCSC-CRL-91-02, University of California, Santa Cruz, 1989.

[8] P. Niyogi and F. Girosi. On the relationship between generalization error, hypothesis complexity, and sample complexity for radial basis functions. *Neural Computation*, 8:819–842, 1996.

[9] I. Koo. Structural optimization of radial basis function networks for function approximation. Master's thesis, Korea Advanced Institute of Science and Technology, 2001.

[10] V. Vapnik. *Statistical Learning Theory.* John Wiley, 1998.

[11] A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, 1965.

[12] G. Lorentz. *Approximation of Functions.* Chelsea Publishing Co., 1986.

[13] A. Barron. Universal approximation bounds for superpositions of a sigmoid functions. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

[14] H. Mhaskar. Approximation properties of a multi-layer feed-forward artificial neural network. *Advances in Computational Mathematics*, 1:61–80, 1993.

[15] F. Girosi and G. Anzellotti. Rate of convergence for radial basis functions and neural networks. In *Artificial Neural Networks for Speech and Vision*, pages 97–113. Chapman and Hall, 1993.

[16] R. M. Kil and I. Koo. Estimation of error confidence intervals for the regression of real-valued functions. In *IEEE International Joint Conference on Neural Networks*, Hawaii, U.S.A., May 2002.