Estimate of error bounds in the improved support vector regression

SUN Yanfeng1, LIANG Yanchun**1,2, WU Chunguo1,2, YANG Xiaowei3,4, LEE Heow Pueh2 and LIN Wu Zhong2

(1. College of Computer Science and Technology, Jilin University, Changchun 130012, China; 2. Institute of High Performance Computing, Singapore 117528, Singapore; 3. Department of Applied Mathematics, South China University of Technology, Guangzhou 510640, China; 4. Centre for ACES, Department of Mechanical Engineering, National University of Singapore, 119260, Singapore)

Received June 3, 2003; revised September 1, 2003

Abstract An estimate of a generalization error bound of the improved support vector regression (SVR) is provided based on our previous work. The boundedness of the error of the improved SVR is proved when the algorithm is applied to the function approximation.

Keywords: support vector regression (SVR), generalization error, approximation error, estimation error.

The study of the error bounds for approximation of an unknown function from scattered, noisy data is becoming more and more popular[1,2] in recent years and has a close relationship with neural networks[3].

In Ref. [4], a novel improved data-dependent support vector regression (SVR) is presented. The kernel function is modified using a conformal mapping to make the kernel data-dependent so as to increase the ability of predicting high noisy data of the method. The implementation process of the improved method consists of the following three steps:

(i) Determine the number of the partitioning points, center and width of the partitions using the optimal partition algorithm (OPA)[5].

(ii) Calculate the basic kernel function using a radial basis function with the following form

\[k(x, x') = \exp(- \|x - x'\|^2 / 2\sigma^2),\]  

where \(\sigma\) is a normalized parameter, \(x\) and \(x'\) are taken from the assigned samples. Modify the basic kernel function using equation

\[\hat{k}(x, x') = D(x)D(x')k(x, x'),\]  

where

\[D(x) = \frac{1}{N}\sum_{i=1}^{N} \exp(- \|x - r_i\|^2 / \tau_i^2),\]  

in which \(N\), \(r_i\) and \(\tau_i\) are the numbers of all the partitioning points, the center and the width of the \(i\)th partition, respectively.

(iii) Train the SVR using the modified kernel function \(\hat{k}(x, x')\).

Simulated results applying the improved method to the prediction of the stock price have demonstrated a better prediction effectiveness and generalization capability than the conventional models. It is worth while to study the improved method theoretically.

Therefore the further study on the improved method is performed on the basis of our previous work[4]. The error bounds of the improved SVR applying to the function approximation from scattered data are examined theoretically. A generalization error is used in the study, which can be decomposed into two parts: an approximation part that is due to the finite number of parameters in the approximation scheme used; and an estimation part that is due to the finite number of data available. We bound each of these two parts under certain assumptions and prove the boundedness of the error for a class of approximation schemes.

1 Decomposition of the generalization error

The function \(f^*(x)\) that has to be approximated is called the target function which has the domain \(X\) and range \(Y\). Let \(X\) and \(Y\) be arbitrary subsets of \(R^n\) and \(R\), respectively. The set \(F\) to which the target function belongs is called the target set. The set \(\hat{H}\) is called approximant set to which all the approximant \(G(x, t_i)\) belongs, where \(t_i\) is parameter of \(G(x, t_i)\). Let there be a probability distribution \(P\)
on the space $X \times Y$ according to which data pairs $(x, y)$ are drawn in independent identical distribution (i.i.d.).

In most cases, the effect of the function approximation can be measured by the generalization error. In Ref. [1], the generalization error is decomposed into two parts: (1) the approximation error that exists due to the finite dimensionality of the manifold to which the approximating functions belongs, and (2) the estimation error that is due to the finiteness of the randomly drawn data. In classical approximation theory[6], an important quantity is the degree of approximation, which depends only on the characteristics of the function being approximated and the class of approximating functions. As the dimensionality of the approximating family (parameterized in some fashion) increases, the approximant converges to the true function. In statistics, one typically assumes the true function belongs to the same family to which the approximating function belongs, but is now known only through a randomly drawn data set. Estimates of this function are constructed from the data and one typically studies the convergence of this estimate to the true function as the data goes to infinity.

Let $f(x)$ be an arbitrary function on $X \rightarrow Y$ and $I[f]$ the expected risk with respect to $P$, then we have the following equation:

$$I[f] = E[(y - f(x))^2]$$

$$= \int_x (y - f(x))^2 dP(x).$$

(4)

Suppose $f_0(x)$ is the solution that minimizes the expected risk in the sense of mean squared error (MSE). It can easily be shown that $f_0 = \arg \min_{f \in F} I[f]$.

In terms of statistical regression, we have

$$f_0(x) = E(y \mid x), \quad E(y \mid x) \in F.$$  

(5)

Let $H_m$ denote a generic subset of $H$ whose elements are parameterized by a number of parameters proportional to $m$. Moreover, assume that the set $H \cap_{i=1}^m$ forms a nested family, that is $H_1 \subseteq H_2 \subseteq \cdots \subseteq H_m \subseteq H$, where functions in $H_m$ have the following form

$$h = \sum_{i=1}^m c_i G(x, t_i),$$

$$G(x, t_i) \in H, i = 1, 2, \cdots, m, \quad \forall h \in H_m.$$  

(6)

Let $f_m(x)$ be the solution of the expected risk $I[f]$ in the sense of MSE, that is

$$f_m = \arg \min_{f \in H_m} I[f].$$  

(7)

In general, what we know is just the samples instead of the distribution function. It is difficult to calculate the expected risk $I[f]$ directly. Therefore, an empirical risk $I_{emp}[f]$ is used here to support the calculation for the expected risk. Suppose $\{(x_i, y_i)\}_{i=1}^n$ is a set of $n$ sample points obtained by sampling $X \times Y$ in i.i.d., then we have

$$I_{emp}[f] = \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2.$$  

(8)

Accordingly, suppose the optimum solution of the empirical risk $I_{emp}[f]$ in the sense of MSE is

$$f_{n, m} = \arg \min_{f \in H_m} I_{emp}[f].$$  

(9)

Summarizing the above and measuring the distance in the $L^2(P)$ metric, we could define the generalization error as

$$\|f_0 - f_{n, m}\|^2 = E[(f_0 - f_{n, m})^2].$$  

(10)

According to the definition of $I[f]$ and using eq. (10), the following equation can be easily obtained

$$\|f_0 - f_{n, m}\|^2 = I(f_{n, m}) - I(f_0).$$  

(11)

The generalization error can be decomposed into two parts: the approximation error $E[(f_0 - f_m)^2] = I(f_m) - I(f_0)$ and the estimation error $I_{emp}[f] - I[f]$. A heuristic understanding for the generalization error can be seen in Fig. 1. Notice that the aim of our problem is to find the optimal solution of the target function $f_0(x)$. But because of the restriction of the number of the parameters in the approximation function, what we can obtain is just the restricted solution $f_m(x)$. The error between $f_0(x)$ and $f_m(x)$ is the approximation error. Furthermore, due to the restriction of the number of the measured samples, what we can obtain is just the restricted solution $f_{n, m}(x)$. The error between $f_m(x)$ and $f_{n, m}(x)$ is the approximation error. In Fig. 1, the length of the segment between points $f_0(x)$ and $f_{n, m}(x)$ is the generalization error.

Fig. 1. Decomposition of the generalization error.
2 Estimate of error bounds in the improved SVR

Lemma 1. [1] Let \( \delta \) be a constant \( (0 < \delta < 1) \). If the empirical risk converges to the expected risk uniformly with probability \( (1 - \delta) \) and a bound of the following form

\[
|I_{emp}[f] - I[f]| \leq \omega(n, m, \delta) \tag{12}
\]

is valid, then the following inequality holds

\[
\|f_n(x) - f_n,m(x)\|^2 \leq 2\omega(n, m, \delta). \tag{13}
\]

Theorem 1. Assume \( H \) is a reproducing kernel Hilbert space (RKHS), \( D(x)D(x_i)k(x, x_i) \) a reproducing kernel in \( H \), \( f_0(x) \) the target function and \( f_0(x) \in H \). Let \( a_i \) and \( a_i^* \) be two kinds of Lagrange multipliers which are denoted using \( a_i^{(*)} \), in which if there exists a \(*\) in the () it represents \( a_i^{*} \), otherwise \( a_i \). If the following inequalities are valid

\[
\|D(x)D(x_i)k(x, x_i)\|^2 \leq d^2,
\]

\[
|a_i - a_i^* - \frac{f_0(x_i)}{d^2}| \leq a, \tag{14}
\]

then we have

\[
\|f_0(x) - \sum_{i=1}^{m}(a_i - a_i^*)D(x)D(x_i)k(x, x_i)\|^2 \\
\leq \|f_0(x)\|^2 + md^2a^2. \tag{15}
\]

Proof.

\[
\|f_0(x) - \sum_{i=1}^{m}(a_i - a_i^*)D(x)D(x_i)k(x, x_i)\|^2 \\
= \|f_0(x)\|^2 \\
+ \sum_{i=1}^{m}(a_i - a_i^*)D(x)D(x_i)k(x, x_i) \\
- 2f_0(x) \sum_{i=1}^{m}(a_i - a_i^*)D(x)D(x_i)k(x, x_i) \\
- \|f_0(x)\|^2 \\
+ \sum_{i=1}^{m}(a_i - a_i^*)f_0(x_i) \\
\leq \|f_0(x)\|^2 \\
+ \sum_{i=1}^{m}(a_i - a_i^*)^2 \|D(x)D(x_i)k(x, x_i)\|^2 \\
- 2m(a_i - a_i^*)f_0(x_i).
\]

This completes the proof.

Combining Lemma 1 with Theorem 1, we have the following theorem:

Theorem 2. If the conditions of Lemma 1 are satisfied and Eq. (14) holds, then we have

\[
\|f_0(x) - f_n,m(x)\|^2 \\
= \|f_0(x)\|^2 \\
+ \sum_{i=1}^{m}(a_i - a_i^*)D(x)D(x_i)k(x, x_i) \\
+ 2m(a_i - a_i^*)f_0(x_i) \\
\leq \|f_0(x)\|^2 + \|f_0(x)\|^2 + md^2a^2.
\]

This completes the proof.

Thus the boundedness of the error of the improved SVR in Ref. [4] is proved.

References