Support Vector Machine and Its Application to System Identification

Shuichi Adachi

Abstract—Support Vector Machines (SVMs) have become a subject of intensive study in statistical learning theory. They have been applied to successfully to classification problems and recently extended to regression problems. Support vector machines for regression problems is called Support Vector Regression (SVR). In this paper, a brief introduction to SVR is presented and then a new system identification method based on SVR is proposed for linear in parameter models. The effectiveness of the proposed method is examined through numerical examples.

Keywords—system identification, statistical learning theory, support vector machines, regularization, robust estimation.

I. INTRODUCTION

SUPPORT Vector Machines (SVM) proposed by V.N. Vapnik early in 90's have become a subject of intensive study in statistical learning theory [1], [2]. They have been applied to successfully to classification problems and recently extended to regression problems. SVM for regression problems is called Support Vector Regression (SVR) [3].

In this paper, a brief introduction to SVR is presented and then a new system identification method based on SVR for linear regression models such as Finite Impulse Response (FIR) and Auto-Regressive with eXtra input (ARX) models [4] is proposed. It will be shown through numerical examples that the proposed method is robust for data with outliers and is effective for identification problem with band-limited input signals. Finally, SVR is applied to actual data which is measured from rolling mill. It will be shown that SVR has a good learning performance for finite data samples.

II. SUPPORT VECTOR REGRESSION

The regression problem considered here is to estimate a vector of parameters of an unknown function $f(\mathbf{x}_i, \mathbf{w})$ by making measurements of the function with error at any point $\mathbf{x}_i$:

$$y_i = f(\mathbf{x}_i, \mathbf{w}) + e_i \quad (1)$$

where the error e is independent of $\mathbf{x}$. In SVR, a function linear in parameters, $\mathbf{w} = [w_1, w_2, \dots, w_m]^T$, that is,

$$f(\mathbf{x}_i, \mathbf{w}) = \sum_{j=1}^m w_j g_j(\mathbf{x}_i) \quad (2)$$

is used to approximate the regression, where $g_j(\mathbf{x})$ is a set of nonlinear transformations chosen *a priori*.

Shuichi Adachi is with the Department of Electrical and Electronic Engineering, Utsunomiya University, 7-1-2 Yoto, Utsunomiya, Tochigi, 321-8585 Japan, E-mail: adachis@cc.utsunomiya-u.ac.jp

A key idea in generalization of SVM from the classification problems to the regression problems is an adoption of a new type of loss function, *i.e.*, so-called ε -insensitive or Vapnik's loss function defined by

$$L_\varepsilon(y_i, f(\mathbf{x}_i, \mathbf{w})) = \begin{cases} 0, & |y_i - f(\mathbf{x}_i, \mathbf{w})| \leq \varepsilon \\ |y_i - f(\mathbf{x}_i, \mathbf{w})| - \varepsilon, & \text{otherwise} \end{cases} \quad (3)$$

where a parameter ε controls the width of the insensitive zone and it determines the number of support vectors.

The loss function is suitable for robust regression [5], *i.e.*, under more relaxed assumptions on the error e_i than the conventional least-squares loss function which is appropriate only for Gaussian noise. If the density of the noise is a symmetric convex function with continuous second derivatives, then minimizing Vapnik's loss function with $\varepsilon = 0$ provides the best mini-max approximation to the regression.

Based on the Structural Risk Minimization (SRM) principle [2], a model structure which provides minimal risk values in

$$S_k = \left\{ \sum_{j=1}^m w_j g_j(\mathbf{x}) : (\mathbf{w} \cdot \mathbf{w}) \leq c_k \right\} \quad (4)$$

is selected where c_k is a constant which is selected *a priori*.

Now, the SVR problem is summarized as follows [6].

SVR Problem (Primal problem): Find the parameter vector $\mathbf{w}$ that minimizes the empirical risk functional

$$R_{emp}(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N L_\varepsilon \left(y_i, \sum_{j=1}^m w_j g_j(\mathbf{x}_i) \right) \quad (5)$$

subject to constraint

$$(\mathbf{w} \cdot \mathbf{w}) \leq c_k \quad (6)$$

given the training data $(\mathbf{x}_i, y_i)$, $i = 1, \dots, N$. $\square$

By introducing slack variables and a regularization constant C , the SVR problem can be transformed to the following dual problem.

Dual Problem: Find the parameters α_i^* and β_i^* , $i = 1, \dots, N$ that maximize the functional

$$Q(\alpha, \beta) = -\varepsilon \sum_{i=1}^N (\alpha_i + \beta_i) + \sum_{i=1}^N y_i (\alpha_i - \beta_i) - \frac{1}{2} \sum_{i,j=1}^N (\alpha_i - \beta_i)(\alpha_j - \beta_j) K(\mathbf{x}_i, \mathbf{x}_j) \quad (7)$$

subject to constraint

$$\sum_{i=1}^N \alpha_i = \sum_{i=1}^N \beta_i, \quad 0 \leq \alpha_i, \beta_i \leq C, \quad i = 1, \dots, N \quad (8)$$

given the training data (x_i, y_i) , $i = 1, \dots, N$, an inner product kernel K , an insensitive zone ε , and a regularization parameter C . $\square$

The dual problem can be solved via standard quadratic programming. Once solutions α_i^* and β_i^* , $i = 1, \dots, N$ are calculated, the regression function is constructed in terms of these values as

$$f(x) = \sum_{i=1}^N \gamma_i^* K(x_i, x) \quad (9)$$

where

$$\gamma_i^* = \alpha_i^* - \beta_i^*. \quad (10)$$

In the regression problem, both ε and C which play an important role for noise and smoothness of the function, respectively, have to be selected *a priori*. However, systematic selection of these parameters has not been proposed.

III. A NEW SYSTEM IDENTIFICATION METHOD BASED ON SVR

In this section, a system identification method based on SVR is proposed for linear regression models such as FIR and ARX models [4]. Since a nonlinear kernel is usually utilized as an inner kernel H in Eq.(7), an explicit expression for parameter w could not be derived. On the other hand, in the proposed method a linear kernel is utilized, so an explicit expression for w can be formulated.

First, an FIR model

$$y_t = \sum_{j=0}^m h_j x_{t-j} + v_t, \quad t = 1, 2, \dots, N \quad (11)$$

is considered where $\{x_t\}$ is the input data, $\{y_t\}$ is the output data, $\{h_j\}$ is the impulse response of the system, and $\{v_t\}$ is white noise. The system identification problem considered here is to estimate $\{h_j\}$ based on input-output data of the system.

In order to calculate the model parameters $\{h_j\}$, a basis function for SVR in Eq.(2) corresponds uniquely to

$$g_j(x_t) = x_{t-j}, \quad j = 0, 1, \dots, m. \quad (12)$$

It is noted that if the purpose of identification is output prediction, other basis functions can be utilized.

Then, the output prediction in the primal problem can be written as

$$\hat{y}_t = \sum_{j=0}^m \hat{h}_j g_j(x_t) = \sum_{j=0}^m \hat{h}_j x_{t-j}. \quad (13)$$

Moreover, it can be expressed in matrix form.

$$\hat{\mathbf{y}}_N = \mathbf{G}_N \hat{\mathbf{h}} \quad (14)$$

where

$$\begin{aligned} \hat{\mathbf{y}}_N &= [\hat{y}_1, \hat{y}_2, \dots, \hat{y}_N]^T, \quad \hat{\mathbf{h}} = [\hat{h}_0, \hat{h}_1, \dots, \hat{h}_m]^T, \\ \mathbf{G}_N &= \begin{bmatrix} g_0(x_1) & g_1(x_1) & \cdots & g_m(x_1) \\ g_0(x_2) & g_1(x_2) & \cdots & g_m(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ g_0(x_N) & g_1(x_N) & \cdots & g_m(x_N) \end{bmatrix} \\ &= \begin{bmatrix} x_1 & x_0 & \cdots & x_{1-m} \\ x_2 & x_1 & \cdots & x_{2-m} \\ \vdots & \vdots & \ddots & \vdots \\ x_N & x_{N-1} & \cdots & x_{N-m} \end{bmatrix}. \end{aligned}$$

On the other hand, the prediction output in the dual problem can be written as

$$\hat{\mathbf{y}}_N = \mathbf{G}_N \mathbf{G}_N^T \gamma^* \quad (15)$$

where a linear kernel

$$K(x_i, x_j) = x_i^T x_j \quad (16)$$

is employed as an inner kernel and $\gamma^* = [\gamma_1^*, \gamma_2^*, \dots, \gamma_N^*]^T$ is calculated by Eq.(10).

From Eqs.(14) and (15), unknown parameter estimates can be calculated by

$$\hat{\mathbf{h}} = \mathbf{G}_N^T \gamma^*. \quad (17)$$

The main purpose of the SVR is to predict output data precisely, so the explicit expression of the parameter estimates such as Eq.(17) has not been derived.

Next, an ARX model which is written as

$$y_t = - \sum_{j=1}^m a_j y_{t-j} + \sum_{j=1}^m b_j x_{t-j} + v_t, \quad t = 1, \dots, N \quad (18)$$

is considered. Since the ARX model is also linear in unknown parameters, $\theta = [a_1, \dots, a_m, b_1, \dots, b_m]^T$, the result for the FIR model can be easily extended, so

$$\hat{\theta} = \mathbf{G}_N^T \gamma^* \quad (19)$$

is derived where

$$\mathbf{G}_N^T = \begin{bmatrix} y_1 & y_2 & \cdots & y_N \\ \vdots & \vdots & & \vdots \\ y_{1-m} & y_{2-m} & \cdots & y_{N-m} \\ x_1 & x_2 & \cdots & x_N \\ \vdots & \vdots & & \vdots \\ x_{1-m} & x_{2-m} & \cdots & x_{N-m} \end{bmatrix}. \quad (20)$$

Moreover, the proposed system identification method can be extended to nonlinear system identification problems as long as the corresponding nonlinear model is linear in parameters such as a discrete-time bilinear model. Details of these results are omitted.

Fig. 1. Comparison between output predictions by the proposed method (solid line) and the conventional LSM (dashed line), where dashdot line represents the measured output.

Fig. 2. Selection of insensitive and regularization parameters.

IV. NUMERICAL EXAMPLES

A. ARX model identification

First, an ARX model identification problem is considered. The parameter vector θ was assumed as

$$\theta = [a_1, a_2, b_1, b_2]^T = [-1.60, 0.680, 0.0364, 0.0437]^T.$$

As an identification input, a PRBS (Pseudo Random Binary Signal) was applied, and the output was corrupted by a white noise $N(0, 0.03^2)$ where number of data was 50.

The output prediction result is shown in Fig.1. The insensitive and regularization parameters were selected by comparing Mean Square Error (MSE) shown in Fig.2, as a result $\varepsilon = 0.022$ and $C = 48.5$ were chosen. Number of the corresponding support vector was 24. The parameter estimates by the proposed method ($\hat{\theta}_{SVR}$) and the conventional Least Squares Method (LSM) ($\hat{\theta}_{LSM}$) were

$$\begin{aligned}\hat{\theta}_{SVR} &= [-1.60, 0.681, 0.0386, 0.0496]^T, \\ \hat{\theta}_{LSM} &= [-1.59, 0.679, 0.0379, 0.0492]^T.\end{aligned}$$

The prediction accuracy and parameter estimates of the proposed method are almost the same as the LSM.

Fig. 3. Input-output data with outlier.

B. FIR model identification with an outlier

An FIR model whose impulse response is given by

$$\{w_j\} = \{1, 0.82, 0.67, 0.55, 0.45, 0.37, 0.3, 0.25, 0.2, 0.17, 0.14, 0.11, 0.09\}$$

is considered. A PRBS was applied as an identification input and the output was corrupted by a white noise $N(0, 0.1^2)$. Moreover, an outlier was contained in output data at 18th point. The input and output data are shown in Fig.3.

The identification results are shown in Figs.4 and 5. In this numerical experiment, $\varepsilon = 0.1$ and $C = 1.4$ were selected by comparing Mean Square Error (MSE). In Fig.4, an output prediction by the proposed method (solid line) and measured output (dashdot line) are compared for the validation data. The output prediction by the LSM is also shown by a dotted line. It is clear that the proposed method has an excellent prediction ability, because it is based on the Vapnik's loss function which is robust against outliers. On the other hand, the LSM has a poor prediction performance due to the outlier.

Fig.5 shows the true impulse response and estimated one. The impulse response of the proposed ($\circ$) and the true one ($+$) are almost the same, however, the impulse response by the LSM ($\square$) fluctuates around the true values. From this example, it is shown that the proposed method is robust against outlier.

C. System identification with band-limited input signal

It is well known that white noise is one of the best input signals for system identification experiments, because it contains a lot of frequency components. However, white noise can not be applied in many practical situations, since it fluctuates stably operated plants. In such cases, system identification with band-limited, that is, smooth input signal, is required. It was reported that the conventional LSM has bad convergence property for the band-limited input signal, so system identification based on regularization was proposed [7].

Fig. 4. Comparison between output predictions by the proposed method (solid line) and the conventional LSM (dashed line), where dashdot line represents the measured output.

Fig. 5. Comparison of impulse responses : true value (+), estimated by SVR (o) and estimated by LSM (*).

Now, the proposed identification method based on SVR has regularization effect, so it is expected that it is effective for band-limited input data. To examine this, the following numerical example is presented.

A band-limited input signal shown in Fig.6 (lower) was employed for system identification. From the power spectrum shown in Fig.7, it is clear that the signal consists of low-frequency components. The system to be identified was the same as the FIR model in the previous example. The output of the system which is shown in Fig.6 (upper) was corrupted by a white noise $N(0, 0.1^2)$ where number of data was 50.

The identification result is shown in Fig.8. In this example, $\varepsilon = 0.04$ and $C = 0.3$ were used for the proposed method. It is clear that the proposed method is effective even when the input is band-limited signal. On the other hand, estimates by the LSM fluctuates seriously around the true values.

Fig. 6. Band-limited input signal (lower) and corresponding output signal (upper).

Fig. 7. Power spectrum of the band-limited signal (dashed line) and white noise (solid line).

V. APPLICATION OF SVR TO ROLLING-MILL DATA

In order to examine the performance of SVR, it is applied to actual data which was measured from a hot strip rolling mill in Japan. The flowchart of the roughing rolling method is shown in Fig.9. The measured data consists of the width of the rolled steel strip after horizontal rolling. It contains 561 instances and each instance includes 7 attributes.

An experimental nonlinear model, which is called a set-up model, for the width of the rolled strip after horizontal rolling (W_O)

$$W_O = W_E + \Delta W_H + \Delta W_D \quad (21)$$

was proposed [8] where

$$\begin{aligned} \Delta W_H &= W_E \left\{ \left(\frac{H_I}{H_O} \right)^{C_H} - 1 \right\}, \\ \Delta W_D &= C_D (W_I - W_E), \\ C_H &= \exp \left\{ b_0 \left(\frac{W_E}{H_I} \right)^{b_1} \cdot \left(\frac{W_E}{L_D} \right)^{b_2 \left(\frac{W_E}{H_I} \right)} \left(\frac{H_I}{R_H} \right)^{b_3 \left(\frac{W_E}{H_I} \right)} \right\}, \\ C_D &= \exp \left\{ a_0 \left(\frac{W_I - W_E}{W_I} \right)^{a_1} \right\} \end{aligned}$$

Fig. 8. Comparison of impulse response : true value ($\times$), estimated by SVR ($\circ$) and estimated by LSM ($\square$).

Fig. 9. Illustration of roughing rolling method.

$$\left\{ \left(\frac{H_I}{R_E} \right)^{a_2} \left(\frac{R_E}{W_I} \right)^{a_3} \left(\frac{W_I}{W_E} \right)^{a_4} \right\}.$$

W_E and W_I are the width of the rolled strip after edging and before rolling, respectively, H_I and H_O are the thickness of the rolled strip before rolling and after horizontal rolling, respectively, and R_E and R_H are the diameter of the edger and horizontal roller, respectively.

The regression problem considered here is to predict W_O based on input variables W_I , W_E , H_I and H_O . To solve the static regression problem Eq.(21), model parameters $a_0 \sim a_4$ and $b_0 \sim b_3$ have to be estimated based on the measured data. However, the model Eq.(21) is very complicated, so satisfactory prediction model could not be obtained by the conventional method.

In this paper, SVR is applied to learn the set-up model. A linear kernel is employed as an inner kernel H and three kinds of candidates for basis

- Case 1 : $\{W_I, W_E, H_I, H_O\}$
- Case 2 : $\left\{W_I, W_E, H_I, H_O, W_E \frac{H_I}{H_O}\right\}$
- Case 3 : $\left\{W_I, W_E, H_I, H_O, W_E \frac{H_I}{H_O}, W_I^2, W_E^2, W_I W_E\right\}$

Fig. 10. Input data: W_I, W_E, H_I, H_O .

Fig. 11. Output data : W_O .

are utilized. For example, a model corresponds to Case 1 is

$$W_O = w_1 W_I + w_2 W_E + w_3 H_I + w_4 H_O$$

which is the most simple regression model, and a model corresponds to Case 3 is

$$W_O = w_1 W_I + w_2 W_E + w_3 H_I + w_4 H_O + w_5 W_E \frac{H_I}{H_O} + w_6 W_I^2 + w_7 W_E^2 + w_8 W_I W_E$$

It is noticed that the basis for SVR such as H_I/H_O is selected based on the experimental model Eq.(21).

Now, SVR is applied to actual data. The input data W_I , W_E , H_I and H_O are shown in Fig.10 and output data W_O is shown in Fig.11 where all data is normalized. For a model building, 50 data from instance 401 to 450 is used.

Table I summarizes the learning results of SVR for Case 1 to 3. Especially, the learning results of Case 3 is shown in Fig.12. It is clear that appropriate selection of basis improves its learning performance. The important point of

TABLE I
COMPARISON OF LEARNING RESULTS FOR VARIOUS BASIS

Case	ϵ	C	No. of SV	MSE	
				SVR	LSM
1	0.02	0.080	48	1.201	1.349
2	0.14	0.010	39	1.029	1.345
3	0.02	0.025	47	0.869	2.423

Fig. 12. Learning result of Case 3.

this example is to select the basis based on the experimental physical insight.

VI. CONCLUSIONS

In this paper, SVR has been briefly introduced and a new system identification method based on SVR has been proposed. It has been shown that the proposed identification method has several advantages, that is, it is robust against outlier and is effective for band-limited input signal through numerical experiments. Also, SVR has been applied to the regression problem of the rolling mill data and it has a good learning performance.

In this paper, linear system identification problems are treated, however, the SVM is expected to be a powerful tool for nonlinear system identification problem [9].

ACKNOWLEDGMENTS

The authors thank Mr. T.Ogawa of Utsunomiya University for his help in the numerical experiments.

REFERENCES

[1] V.N.Vapnik, *Statistical Learning Theory*, J. Willey & Sons, 1998.
 [2] V.N.Vapnik, *The Nature of Statistical Learning Theory (2nd Edition)*, Springer-Verlag, 1999.
 [3] B. Schölkopf, et al. ed., *Advances in Kernel Methods*, The MIT Press, 1999.
 [4] L. Ljung, *System Identification – Theory for the User (2nd Edition)*, Prentice Hall PTR, 1999.

[5] P.J.Huber, *Robust Statistics*, Wiley, New York, 1981.
 [6] A.J.Smola and B. Schölkopf, *A tutorial on support vector regression*, NeuroCOLT2 Technical Report Series, NC2-TR-1998-030, ESOPRIT Working Group in Neural and Copmputational Learning II, 1999.
 [7] A.Sano, S.Adachi and H.Ohmori, *Convergency improvement in least squares type of adaptive digital filter*, Signal Processing IV : Theories and Applications, edited by J.L. Lacoume, pp.815–818, Elsevier Science Publishers B.V., North-Holland, 1988.
 [8] T.Shibata,, et al., *Edger set-up model at roughing train in hot strip mill*, Iron and Steel, Vol.15, pp.2509–2514, 1981.
 [9] J.A.K.Suykens, *Support vector machines: A nonlinear modelling and control perspective*, European Journal of Control, Vol.7, pp.311–327, 2001.