Online Support Vector Machine with Adaptive Kernel Functions

Hua Zheng[1,a,*], Dongzhu Zhao[1,b], Yafei Shang[1,c] and Shiqiang Duan[1,d]

[1]School of Power and Energy, Northwestern Polytechnical University

Shaan Xi, Xi'an 710072, China

[a]isea_zh@mail.nwpu.edu.cn; [b]dongzhuzhao@mail.nwpu.edu.cn;

[c]yafei@nwpu.edu.cn; [d]duanshiqiang@mail.nwpu.edu.cn

*corresponding author

Keywords: Adaptive Kernel Function; Truncation error minimization criterion; Structural risk minimization; Least Square Support Vector Machine

Abstract. Based on the structural risk minimization criterion, an online Support Vector Machine (SVM) algorithm with adaptively selecting kernel functions is presented. In order to overcome the problem that the traditional method converges slowly and cannot adaptively select samples, this paper first sets the appropriate window function for the time series signal, and then selects the appropriate truncation error minimization criterion of the Least Square Support Vector Machine (LS-SVM) in the process of sample update. The Lagrange factor finally completes the retraining of the new samples. Compared with the traditional SVM method, the results of numerical simulation show that, the proposed algorithm has the characteristics of high prediction accuracy and strong generalization ability, and could be widely used in a series of engineering applications such as pattern recognition, fault diagnosis, machine vision and intelligent control.

1. Introduction

Support Vector Machine (SVM) as a classical supervised learning algorithm can perform classification and regression operations, especially in the analysis of small samples, as well as high-dimensional nonlinear problems, which has obvious advantages[1][2]. The basic idea of a support vector machine is:

a) Support vector machine classify the sample with the largest interval, which use the limited sample information to find the balance between the learning ability of the model and the complexity of the model, based on the structural risk minimization strategy, so as to obtain the strong generalization ability.

b) The learning strategy of support vector machine is interval maximization. It essentially solves the problem of convex quadratic optimization. Theoretically, it could obtain the global optimal solution and solve the local optimal problem that is often difficult for the Back-Propagation neural network.

c) For nonlinear problems, SVM can map the indivisible samples in low-dimensional space to the high-dimensional space through kernel functions, so as to construct linear decision models to make them linearly separable in high-dimensional space.

Unlike conventional SVM, since the equality constraint in Least Square Support the Vector Machine (LS-SVM) replace the inequality constraint, the deviation caused by the empirical risk is

  DOI: 10.25236/ciais.2019.029

changed from the primary to the quadratic. This greatly reduces the computational complexity of the algorithm and improves the computational speed of solving the problem [3]. From the point of view in mathematical analysis, although the conversion between the inequality constraints and the equality constraints lack of adequate theoretical basis, the advantages of its simplicity and practicability attract many scholars in engineering research field.

The kernel function (nonlinear function) has achieved good results under the condition of data batch processing, however, how to implement the kernel function algorithm online is a core issue in practical applications [4][5]. The main factors affecting the online implementation of existing kernel function algorithms include: (1) Because the dimension of the weight coefficient in Hilbert space is too high, it is easy to overfitting. (2) In the traditional kernel function-based estimator, the function expression becomes more complicated as the number of observed samples grows. (3) The training time of the data batch or data incremental update algorithm will increase linearly with the number of observed samples. For the above reasons, a simple online regular risk minimization algorithm based on stochastic gradient descent had been proposed (Naive Online Minimization Algorithm, NORMA), which can not only realize the kernel function target online, but also map the input sample to the high-dimensional space, and the convergence speed of the algorithm can still meet the requirements, and the error limit can be effectively analyzed.

Although the NORMA algorithm can achieve effective tracking, since the implementation of the algorithm is based on truncating samples to achieve acceleration, it can hardly selectively add or delete samples, and. In practical applications, the learning algorithm that adaptively selects the kernel function can really meet the computational requirements and also meet the requirements of the regular minimization algorithm. Based on this, the use of adaptive kernel function principle is proposed and studied simple adaptive regularization risk minimization algorithm (Adaptive Minimization Naive Online Mathimatics-Numerical algorithms, ANORMA).

2.Adaptive Kernel Functions

2.1 Background. In the dual space, the solution of the LS-SVM depends on all training samples. Given a training sample $\{x_i, y_i\}_{i=1}^{N}$. When the training samples arrive in sequence, the objective function is as followed[6][7]:

$$J(\omega, b) = \frac{1}{2} \| \omega \|^2 + \gamma \sum_{k=1}^{N} (y_k - \omega \phi(x_k) - b)^2 \tag{1}$$

In the Eq.1, ω is a weight vector, γ is a regularization parameter, which can comprehensively consider the training error and model complexity, so that the function sought has good generalization ability. According to statistical learning theory, the solution of this form of support vector machine is

$$y(x) = \sum_{k=1}^{N} a_k k(x_k, x) + b \tag{2}$$

where, a_k is a support vector and $k(\ )$ is a kernel function.

2.2 Proposed Algorithm. First, the online method is used to train the parameters of the initial data, the new data samples obtained later are added to the training set, the data with small

contribution is deleted, and the data set after deleted is used as a new training set, and the parameters are recalculated. One advantage of LS-SVM is that it is easier to use iterative update to remove and add training samples.

Setting a given sample set $\{x_k, y_k\}_{k=1}^{N}, x_k \in \boldsymbol{R}^d, y_k \in \boldsymbol{R}$, the function based on SVM is:

$$y(x) = \omega\phi(x) + b \tag{3}$$

where, ω is the weight vector and b is the deviation. $\phi(\ )$ map the input space to a high-dimensional (possibly infinite-dimensional) feature space, a kernel function can be used to release this step. The kernel function can be implemented using functions in the original space, and does not require a specific form of $\phi(\ )$. The optimal ω and b can be obtained by minimizing the objective function consists the sum of the fitting errors and the regularization term. The objective function used is based on the adaptive selection of samples, and the adaptive simple regular risk function algorithm needs to be calculated is （ f_t， $\mathrm{R}_{inst,\lambda}[f, x_t, y_t]|_{f=f_t}, a_t, b_t$） . Where $f_t(x)$ is the estimate to $y(x)$,and $\mathrm{R}_{inst,\lambda}[f, x_t, y_t]|_{f=f_t}$ is the instantaneous risk function.

According to the representation method of the above support vector machine, $f_t(x)$ can be written to $f_t(x) = \sum_{i=1}^{t-1} a_i k(x_i, x) + b$, $x \in \mathrm{X}$,the traditional least squares support vector machine could use the KKT (Karuslr Kuhn Tucker) condition to find the optimal $(a, b)^T$:

$$\begin{pmatrix} \Psi & \Theta^T \\ \Theta & \boldsymbol{\Omega} \end{pmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 0 \\ Y \end{bmatrix} \tag{4}$$

where $\Psi = [0, \cdots, 0]; \Omega = (\boldsymbol{K} + \gamma^{-1}\mathbf{I}), \quad \boldsymbol{K} = \{k_{\mathbf{ki}} = k(x_{\mathbf{k}}, x_{\mathbf{i}})\}_{\mathbf{k,i=1}}^{\mathbf{N}}; \Theta = [1, \cdots, 1]^{\mathbf{T}}; Y = [y_1, \cdots, y_{\mathbf{N}}]^{\mathbf{T}}$,

$a = [a_1, \cdots, a_N]^T, k(\ )$ is a kernel function, $\mathbf{I}$ is the unit matrix. However, this method has limitations in implementing iterative update of weights, so it is more desirable to get updated recursive function expressions.

The recursive function expression $f_{t+1} := f_t - \eta_t \partial_f R_{inst,\lambda}[f, x_t, y_t]|_{f=f_t}, i \in N, f_i \in H, \partial_f$ is the abbreviation of $\partial / \partial f$ (corresponding to the gradient of f), $\eta_t > 0$ is the learning rate, which is often a constant.

Determine the instantaneous risk function $\mathrm{R}_{inst,\lambda}[f, x_t, y_t]|_{f=f_t}$. Assuming f is a linear model, according to the Empirical risk minimization (ERM) criteria, the instantaneous risk error could be defined as ($\mathrm{R}_{inst,\lambda}[f, x, y] := \mathrm{R}_{emp}[f] + \frac{\lambda}{2} \| f \|_{\mathrm{H}}^2$). An important point to mention is that the SVM

optimization problem can be understood as a regularization problem.

$$\min_{f,b} \frac{1}{\mu} \sum_{i=1}^{n} \mathrm{L}(y_i(f(x_i)+b)) + \lambda \| f \|_H^2 \tag{5}$$

where $f(\mathbf{x}) = \boldsymbol{\omega}^T \phi(\mathbf{x})$; $\| f \| := \| \omega \|$; L is a penalty function. If selected $\mathrm{L}(z) = (1-z)^2$, the regularized risk error expression of the least squares support vector base can be obtained. In the regression process discussed in this paper, the penalty function is used. $l(f(x), y) := \frac{1}{2}(y - f(x))^2$

The update factor can be written as: $(a_t := -\eta\, l'(f_t(x_t), y_t) = \eta\, (y_t - f(x_t), i = t, a_i = (1-\eta_t \lambda) a_i;$

$b_{t+1} = b_t - \eta \partial_b R_{inst}[g, x_t, y_t]|_{g=f_t+b_t} = b_t + \eta(y_t - f(x_t))), \mathrm{i<t};$)

Choosing to delete the smallest amplitude a_i does not guarantee a minimum error change, so here proposed an improved algorithm. Relatively selecting the Lagrange factor with the smallest amplitude, the improved algorithm selects the one of $a_i / A(i,i)^{-1}$ with the smallest amplitude to delete. Among them,

$$A(k) = \begin{pmatrix} 0 & \Theta^T \\ \Theta & \boldsymbol{U(k)} \end{pmatrix} \quad \Theta = [1, \cdots, 1]^T; \boldsymbol{U(k)} = (\boldsymbol{K}(k) + C^{-1}\mathbf{I}) \tag{6}$$

where, $\boldsymbol{K} = \{k_{ki} = k(x_k, x_i)\}_{k,i=1}^{N}$; $\boldsymbol{K}(k)$ Represents the matrix calculated at the k moment.

3. Simulation and Analysis

Without lossing generality, select the model $y = \sin(x)/x$, defined $t = -5$ to $t = 5$ within a range of randomly selected points, additional zero mean and a standard deviation of 0.1 Gaussian white noise, the kernel function is used Gaussian kernel function $k(x, x') = \exp(-(x-x')^T (x-x') / (2\sigma^2))$, taking $\sigma = 2.0$.

3.1 Accuracy. The simulation experiment was performed based on a random sample of additional zero-mean Gaussian white noise. Firstly, consider the effects of the two algorithms on the convergence accuracy, and the length of window are taken respectively $N = 20$ and $N = 80$. Fig. 1 and Fig. 2 are the convergence curves of the lumped mean square error.

The green line in Fig. 1 and Fig. 2 is the learning curve of the improved algorithm, and the red line is the learning curve of the original algorithm. Fig. 1 shows the comparison of the original signal plus a Gaussian white noise with a standard deviation of 0.1 to the original NORMA algorithm, and it's size is $N = 20$. The corresponding size in Fig. 2 is $N = 80$. As can be seen from the figures, even the window length and the signal-to-noise ratio are the same, the ANORMA convergence rate is almost the same, but a smaller square error can be obtained.

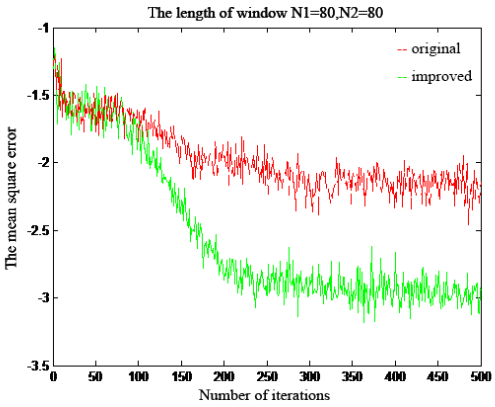

Figure 1. Window length N=20

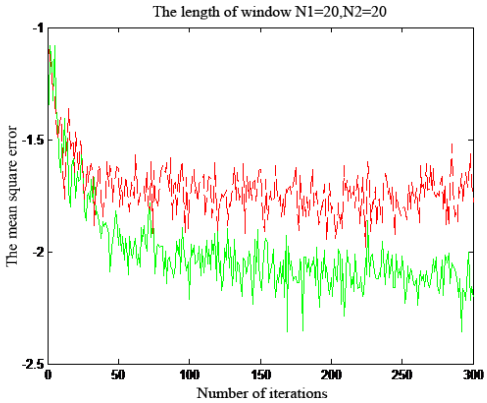

Figure 2. Window length N=80

3.2 SNR Influence. In order to verify the reliability of the algorithm under different Signal To Noise Ratio (SNR) conditions, Gaussian white noise with a zero mean standard deviation of 0.2 is re-added.

The Fig. 3 is a comparison of the original signal plus a Gaussian white noise with a standard deviation of 0.2, ANORMA to the original NORMA algorithm. The window's size is $N = 20$. The corresponding window's size in Fig. 4 is $N = 80$. It can be seen from the figure that in the case of the same window length, although the signal-to-noise ratio increases and the ANORMA convergence rate is almost the same, a smaller convergence result can be obtained.

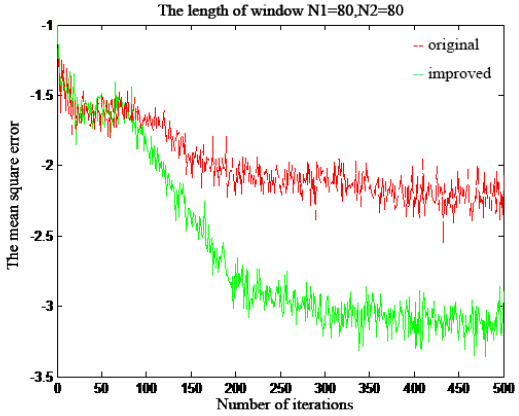

Figure 3. Window length N=20

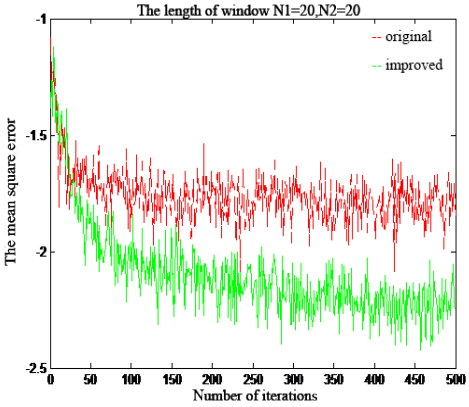

Figure 4. Window length N=80

3.3 Discussion Of Window Length. To further verify the feasibility of the new algorithm, the two algorithms take different window lengths. Fig. 5 takes the ANORMA window's length to 20, the NORMA window's length to 80, and draws the lumped average convergence curve; Fig. 6 takes the ANORMA window's length to be 80 and the NORMA window's length to 20, drawing a lumped average convergence curve.

As can be seen from Fig. 5, the ANORMA algorithm still achieves similar convergence accuracy when the ANORMA algorithm window length is 20 and the NORMA algorithm window length is 80. The ANORMA algorithm's convergence accuracy is greatly improved when the ANORMA algorithm window is 80, and the NORMA algorithm window length is 20.

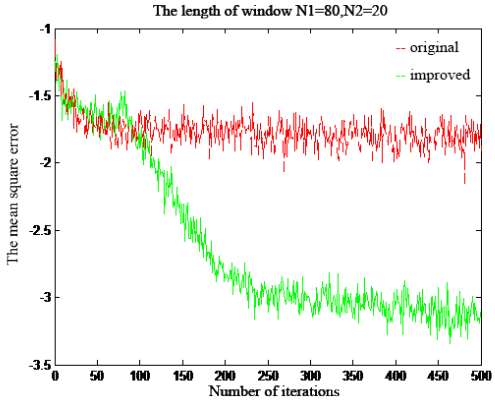

Figure. 5 window length N1=20, N2=80

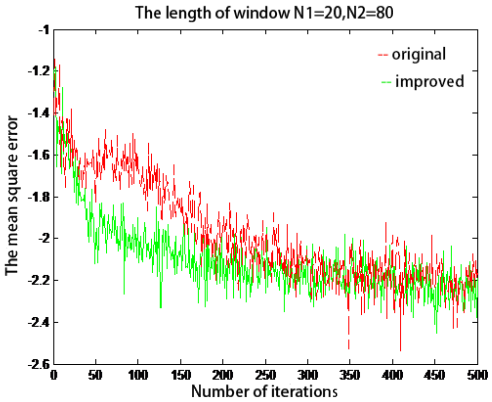

Figure. 6 window length N1=80, N2=20

4.Conclusion

This paper studies an adaptive online algorithm based on LS-SVM for regression prediction. In the process of using LS-SVM to predict, the adjacent data samples are no longer used. The improved algorithm selects the samples which contribute the most to the prediction, and the other samples are deleted. This step complete improvements to the original NORMA algorithm. The simulation results show that the proposed algorithm not only has good generalization ability, but also has smaller prediction error, which can meet the requirements of real-time data processing and prediction in actual engineering.

Acknowledgements

This work is supported by the Fundamental Research Funds for the Central Universities (Grant no. 31020190MS702) and National Science and Technology Major Project(2017-V-0011-0062).

References

[1] Deng W, Yao R, Zhao H, et al. A novel intelligent diagnosis method using optimal LS-SVM with improved PSO algorithm[J]. Soft Computing, 2019, 23(7): 2445-2462.

[2] Zheng H, Zhang Y, Liu J, et al. A novel model based on wavelet LS-SVM integrated improved PSO algorithm for forecasting of dissolved gas contents in power transformers[J]. Electric Power Systems Research, 2018, 155: 196-205.

[3] Asfaram A, Ghaedi M, Azqhandi M H A, et al. Statistical experimental design, least squares-support vector machine (LS-SVM) and artificial neural network (ANN) methods for modeling the facilitated adsorption of methylene blue dye[J]. Rsc Advances, 2016, 6(46): 40502-40516.

[4] Langone R, Alzate C, De Ketelaere B, et al. LS-SVM based spectral clustering and regression for predicting maintenance of industrial machines[J]. Engineering Applications of Artificial Intelligence, 2015, 37: 268-278.

[5] Roxas E A, Vicerra R R P, Lim L A G, et al. SVM Compound Kernel Functions for Vehicle Target Classification[J]. Journal of Advanced Computational Intelligence and Intelligent Informatics, 2018, 22(5): 654-659.

[6] Chen T T, Lee S J. A weighted LS-SVM based learning system for time series forecasting[J]. Information Sciences, 2015, 299: 99-116.
[7] Ji J, Zhang C, Gui Y, et al. New observations on the application of LS-SVM in slope system reliability analysis[J]. Journal of Computing in Civil Engineering, 2016, 31(2): 06016002.