Training for Huge Data set with On Line Pruning Regression by LS-SVM

Daehak Kim[1], Jooyong Shim[2] and Kwangsik Oh[3]

Abstract

LS-SVM(least squares support vector machine) is a widely applicable and useful machine learning technique for classification and regression analysis. LS-SVM can be a good substitute for statistical method but computational difficulties are still remained to operate the inversion of matrix of huge data set. In modern information society, we can easily get huge data sets by on line or batch mode. For these kind of huge data sets, we suggest an on line pruning regression method by LS-SVM. With relatively small number of pruned support vectors, we can have almost same performance as regression with full data set.

Key-words : On line Regression, Pruning, LS-SVM, Huge data set, Training.

1. Introduction

The least squares support vector machine(LS-SVM), a modified version of support vector machine introduced by Vapnik(1995, 1998) in a least squares sense, has been proposed for classification and regression by Suykens and Vanderwalle(1999). In LS-SVM the solution is given by a linear system instead of a quadratic program problem. The fact that LS-SVM has an explicit primal-dual formulations has a number of advantages. However a drawback of LS-SVM is that the sparseness is lost differently from in Vapnik's SVM. Suykens *et al.*(2000) suggested a procedure imposing the sparseness by gradually pruning the support vectors based on sorted absolute values of optimal lagrange multipliers which result from the solutions to the linear system of LS-SVM. And they illustrate that support vectors can be decreased from 500 to less than 100 without loss of performance in the sine example.

But the LS-SVM algorithms are trained in batch form, which is not suited to the real application such as on line system and control, where the data come in sequentially or the size of data is huge. So the on line training for the regression is needed urgently in real application.

We suggest an on line pruning regression method by LS-SVM which always uses the fixed number of support vectors and their corresponding outputs being modified at each time of a new data point coming in. These modified pruned support vectors and their corresponding outputs are used to predict the regression function of the testing data set.

1) Professor, Dept. of Statistical Information, Catholic University of Daegu, Kyungbuk, 712-702
2) Adjunct Professor, Dept. of Statistical Information, Catholic University of Daegu.
3) Professor, Dept. of Statistical Information, Catholic University of Daegu, Kyungbuk, 712-702

The performance of the proposed method is almost same with that of the method of Suykens *et al.*(2000), and not much inferior to that of the LS-SVM based on full data set.

2. LS-SVM Regression

Let the training data set be denoted by $\{ \boldsymbol{x}_i, y_i \}_{i=1}^N$, with each input $\boldsymbol{x}_i \in R^d$ and the output y_i which is the output corresponding to $\boldsymbol{x}_i$. The LS-SVM regression takes the form

$$f(\boldsymbol{x}) = \boldsymbol{w}'\phi(\boldsymbol{x}) + b$$

where the term b is a bias term. Here the feature mapping function $\phi(\cdot): R^d \rightarrow R^{d_f}$ maps the input space to the higher dimensional feature space where the dimension d_f is defined in an implicit way. The optimization problem is defined with a regularization parameter C as

$$\text{Minimize } \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + \frac{C}{2}\sum_{i=1}^{N} e_i^2 \tag{1}$$

over $\{\boldsymbol{w}, b, \boldsymbol{e}\}$ subject to equality constraints

$$y_i = \boldsymbol{w}'\phi(\boldsymbol{x}_i) + b + e_i, \quad i = 1, \cdots, N.$$

The Lagrangian function can be constructed as

$$L(\boldsymbol{w}, b, e; \alpha) = \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + \frac{C}{2}\sum_{i=1}^{N} e_i^2 - \sum_{i=1}^{N} \alpha_i (\boldsymbol{w}'\phi(\boldsymbol{x}_i) + b + e_i - y_i) \tag{2}$$

where α_i's are the Lagrange multipliers. The Karush-Kuhn-Tucker(Smola and Scholkopf, 1998) conditions for optimality are given by

$$\frac{\partial L}{\partial \boldsymbol{w}} = 0 \rightarrow \boldsymbol{w} = \sum_{i=1}^{N} \alpha_i \phi(\boldsymbol{x}_i)$$

$$\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{N} \alpha_i = 0$$

$$\frac{\partial L}{\partial e_i} = 0 \rightarrow \alpha_i = C e_i, \quad i = 1, \cdots, N$$

$$\frac{\partial L}{\partial \alpha_i} = 0 \rightarrow \boldsymbol{w}'\phi(\boldsymbol{x}_i) + b + e_i - y_i = 0, \quad i = 1, \cdots, N,$$

with solution

$$\begin{bmatrix} 0 & \mathbf{1}' \\ \mathbf{1} & \Omega + C^{-1} I \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{y} \end{bmatrix} \tag{3}$$

with $\boldsymbol{y} = (y_1, \cdots, y_N)'$, $\mathbf{1} = (1, \cdots, 1)'$, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_N)'$, and $\Omega = \{\Omega_{kl}\}$ where $\Omega_{kl} = \phi(\boldsymbol{x}_k)'\phi(\boldsymbol{x}_l) = K(\boldsymbol{x}_k, \boldsymbol{x}_l)$, $k, l = 1, \cdots, N$, which are obtained from the application of Mercer's conditions(1909). Several choices of the kernel $K(\cdot, \cdot)$ are possible.

By solving the linear equation (3) the optimal bias and Lagrange multipliers, $\hat{b}$ and $\hat{\alpha}_i$'s can be obtained, then the optimal target value for the given $\boldsymbol{x}$ is obtained as

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{N} \hat{\alpha}_i K(\boldsymbol{x}, \boldsymbol{x}_i) + \hat{b}. \quad (4)$$

3. On line Pruning Regression by LS-SVM

The loss of sparseness of LS-SVM follows from the fact that Lgrange multipliers are proportional to the error at the data point. However, by sorting absolute values of α_i one can evaluate which data are most significant for the contribution to LS-SVM model. Suyens *et al.*(2000) imposed the sparseness by gradually eliminating the least significant data points from the training data set and reestimating LS-SVM by the following procedures.

(a) Train LS-SVM based on N data points.
(b) Remove a small amount of data points(e.g. 5%) with smallest absolute values of α_i's.
(c) Retrain LS-SVM based on the reduced training data set.
(d) Iterate (b) and (c) until the performance degrades.

Since the pruning method above operate in a batch mode, the inversion of matrix in the linear system of LS-SVM for training the huge data set is computationally difficult.

Now we propose a solution to train the huge data set for LS-SVM. Starting with the fixed number, Nw, of data points of the training data set, we can obtain the optimal Lagrange multipliers and bias by solving the linear systems of LS-SVM. Then we can predict the regression function of the testing data set with them. When the next data point comes, we obtain the optimal Lagrange multipliers by solving the linear system of LS-SVM based on $\{\{\boldsymbol{x}_i, y_i\}_{i=1}^{Nw}, (\boldsymbol{x}_{Nw+1}, y_{Nw+1})\}$. Next by eliminating the data point with smallest absolute value of α_i, we have Nw pruned support vectors and their corresponding outputs, $\{\boldsymbol{x}_i^*, y_i^*\}_{i=1}^{Nw}$, which can be used to predict the regression function of the testing data set via LS-SVM.

Consider that we have Nw pruned support vectors and corresponding outputs based on the first n data points of the training data set and that now the new data points $(\boldsymbol{x}_{n+1}, y_{n+1})$ is coming in. Then we obtain the optimal Lagrange multipliers by solving the linear system of LS-SVM based on $\{\{\boldsymbol{x}_i^*, y_i^*\}_{i=1}^{Nw}, (\boldsymbol{x}_{n+1}, y_{n+1})\}$, where $\{\boldsymbol{x}_i^*, y_i^*\}_{i=1}^{Nw}$ is the pruned support vectors and their corresponding outputs selected from n data points of the training data set. Next by eliminating the data point with smallest absolute value of α_i, we have new Nw pruned support vectors and their corresponding outputs, $\{\boldsymbol{x}_i^*, y_i^*\}_{i=1}^{Nw}$, which can be used to predict the regression

function of the testing data set via LS-SVM.

4. Numerical Study

We illustrate the performance of the proposed algorithm through the simulated data. For the nonlinear regression model the response variables y_i's can be expressed as

$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i, \quad i = 1, \cdots, N,$$

where ε_i is assumed to have zero mean and finite variance. For the generation of data, we set the true value of the regression function to $f(x) = \sin(x) + 0.5$ given the covariate x. For training data set, 1000 of x's are generated from a uniform distribution, $U(-\pi, +\pi)$, and 1000 of ε's are generated from normal distribution, $N(0, 0.3^2)$. For the test data set, 1000 of $(\boldsymbol{x}, \varepsilon)$'s are generated by the same way as for the training data set. The radial basis kernel function is employed for the nonlinear regression, which is

$$K(\boldsymbol{x}_k, \boldsymbol{x}_l) = \exp\left(-\frac{\| \boldsymbol{x}_k - \boldsymbol{x}_l \|^2}{\sigma^2}\right).$$

For the optimization problem in (1), the value of regularization parameter C is chosen as 100 and the bandwidth parameter in the radial basis kernel function σ is chosen as 1. With the fixed number of pruned support vectors as 200, we modify pruned support vectors at each time of new data point coming in.

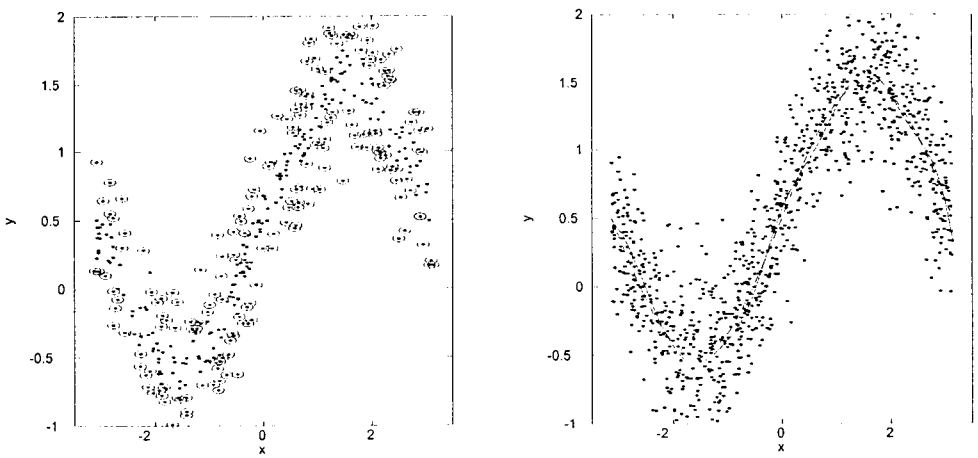

Figure 1. The scatter plot of 200 pruned support vectors of 400 training data and estimated regression function for 1000 testing data.

Figure 1 shows the scatter plot of outputs versus support values of 400 training data. 400 data points are denoted by "·" and those by "o" are 200 on line pruned support values. In the figure 1(right), estimated regression function based on pruned support vectors and the testing data set of 1000 data points is given. Solid line is the LS-SVM estimator

of regression function based on 200 on line pruned support values selected from 400 training data points and dashed line is the true regression function. Figure 2 represents the same results as in Figure 1 but from 1000 training data. From the figures, we can note the estimated function based on 200 on line pruned support values is quite similar to the true function.

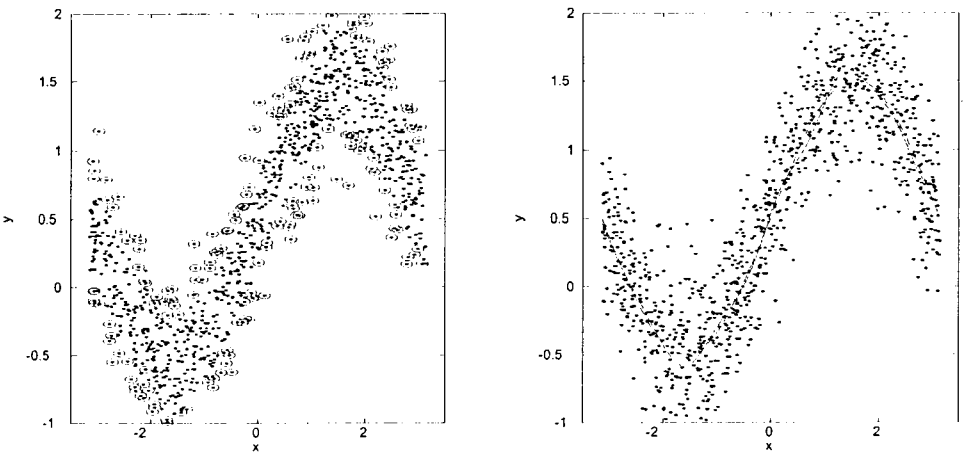

Figure 2. The scatter plot of 200 pruned support vectors of 1000 training data and estimated regression function for 1000 testing data.

5. Concluding Remarks

In this paper, we suggested an on line pruning regression method by LS-SVM. By using relatively small number of pruned support vectors, we can have almost same performance as regression with full data set. Proposed method can be applied to the analysis of on line system and control where the data come in sequentially or the size of data is huge.

References

Mercer, J. (1909). Functions of Positive and Negative Type and Their Connection with Theory of Integral Equations. *Philosophical Transactions of Royal Society,* A, 415-446.

Smola, A. and Scholkopf, B. (1998). On a Kernel-Based Method for Pattern Recognition, Regression, Approximation and Operator Inversion. *Algorithmica*, 22, 211-231.

Suyken, J.A.K., Lukas L., and Vandewalle J. (2000). Sparse Approximation using Least Squares Support Vector Machines, *IEEE International Symposium on Circuits and Systems*(ISCAS 2000). 757-760, Geneva, Switzerland.

Suykens, J.A.K. and Vanderwalle, J. (1999). Least Square Support Vector Machine Classifier, *Neural Processing Letters*, 9, 293-300.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory.* Springer, New York.

Vapnik, V. N. (1998). *Statistical Learning Theory*, Springer, New York.