Journal of Korean
Data & Information Science Society
2005, Vol. 16, No. 4, pp. 1087~1094

e-SVR using IRWLS Procedure

Jooyong Shim[1)]

Abstract

e-insensitive support vector regression(e-SVR) is capable of providing more complete description of the linear and nonlinear relationships among random variables. In this paper we propose an iterative reweighted least squares(IRWLS) procedure to solve the quadratic problem of e-SVR with a modified loss function. Furthermore, we introduce the generalized approximate cross validation function to select the hyperparameters which affect the performance of e-SVR. Experimental results are then presented which illustrate the performance of the IRWLS procedure for e-SVR.

Keywords : e-insensitive support vector regression, Generalized approximate cross validation function, Iterative reweighted least squares procedure, Kernel function

1. Introduction

SVM, firstly developed by Vapnik(1995, 1998), is being used as a new technique for regression and classification problems. SVM is based on the structural risk minimization(SRM) principle, which has been shown to be superior to traditional empirical risk minimization(ERM) principle. SRM minimizes an upper bound on the expected risk unlike ERM minimizing the error on the training data. By minimizing this bound, high generalization performance can be achieved. In particular, for the SVM regression case SRM results in the regularized ERM with the e-insensitive loss function. The introductions and overviews of recent developments of SVM regression can be found in Vapnik(1995,1998), Smola and Scholkopf(1998), and Wang(2005).

Training an SVR requires the solution to a quadratic programming(QP) optimization problem. But QP problem presents some inherent limitations which

1) Department of Applied Statistics, Catholic University of Daegu, Kyungbuk, 712-702, Korea.
E-mail : ds1631@hanmail.net

results in computational difficulty especially for the large data sets. Platt(1998) developed the sequential minimal optimization(SMO) algorithm which divides the QP problem into a series of small QP problems to avoid such computational difficulty. Perez-Cruz et al.(2000) proposed IRWLS algorithm for SVR by transforming the Lagrangian function into sum of quadratic terms by defining associated weights of predicted errors.

In this paper we propose an IRWLS procedure to solve the QP problem of e-SVR with a modified loss function of which original version is e-insensitive loss function used by Vapnik(1995, 1998). The modified loss function is attained by providing the differentiability at $\pm e$, which enables to solve QP problem by IRWLS procedure. To select appropriate hyperparameters, a commonly used method is minimizing the cross validation(CV) function. Nychka et al.(1995) proposed the approximate cross validation(ACV) function for quantile spline estimation. This technique can be easily applied to e-SVR using IRWLS. And by replacing each element of hat matrix by the average of trace of hat matix, the GACV function also can be obtained. GACV function is used to select hyperparameters for the achievement of high generalization performance. The rest of this paper is organized as follows. In Section 2 we give a review of e-SVR. In Section 3 we propose an IRWLS procedure for e-SVR and present the model selection method using GACV function. In Section 4 we perform the numerical studies through examples. In Section 5 we give the conclusions.

2. e-Support Vector Regression

Let the training data set D be denoted by $(\boldsymbol{x}_i, y_i)_{i=1}^n$, with each input $\boldsymbol{x}_i \in R^d$ including a constant 1 and the response $y_i \in R$, where the output variable y_i is related to the input vector $\boldsymbol{x}_i$. 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. An inner product in feature space has an equivalent kernel in input space, $\phi(\boldsymbol{x}_i)'\phi(\boldsymbol{x}_j) = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ (Mercer(1909)). Several choices of the kernel $K(\cdot, \cdot)$ are possible. We consider the nonlinear regression case, in which the regression function of the response given $\boldsymbol{x}$, $f(\boldsymbol{x})$, can be regarded as a nonlinear function of input vector$\boldsymbol{x}$.

With e-insensitive loss function $\rho_e(\cdot)$, the estimator of the regression function can be defined as any solution to the optimization problem,

$$\min \frac{1}{2}\boldsymbol{w}'\boldsymbol{w} + C\sum_{i=1}^{n} \rho_e(y_i - f(\boldsymbol{x}_i)), \tag{1}$$

where $\rho_e(r)=0$ if $|r| \leqq e$ and $\rho_e(r)=r-e$ if $|r| > e$. We can express the regression problem by formulation for SVR as follows.

$$\min \frac{1}{2}\boldsymbol{w}'\boldsymbol{w}+ C\sum_{i=1}^{n} (\xi_i+\xi_i^*) \tag{2}$$

subject to

$$\begin{aligned} y_i-\boldsymbol{w}'\phi(\boldsymbol{x}_i) &\leqq e+\xi_i \\ \boldsymbol{w}'\phi(\boldsymbol{x}_i)-y_i &\leqq e+\xi_i^* \\ e, \xi_i, \xi_i^* &\geqq 0 \end{aligned}$$

where C is a regularization parameter penalizing the training errors.

We construct a Lagrange function as follows:

$$\begin{aligned} L = \frac{1}{2}\boldsymbol{w}'\boldsymbol{w}+ C\sum_{i=1}^{n} (\xi_i+\xi_i^*) - \sum_{i=1}^{n} \alpha_i(e+\xi_i-y_i+\boldsymbol{w}'\phi(\boldsymbol{x}_i)) \\ - \sum_{i=1}^{n} \alpha_i^*(e+\xi_i^*+y_i-\boldsymbol{w}'\phi(\boldsymbol{x}_i)) - \sum_{i=1}^{n} (\eta_i\xi_i+\eta_i^*\xi_i^*). \end{aligned} \tag{3}$$

We notice that the positivity constraints $\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \geqq 0$ should be satisfied. After taking partial derivatives of equation (3) with regard to the primal variables $(\boldsymbol{w}, \xi_i, \xi_i^*)$ and plugging them into equation (3), we have the optimization problem below.

$$\max -\frac{1}{2}\sum_{i,j=1}^{n} (\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_{i=1}^{n} (\alpha_i-\alpha_i^*)y_i - e\sum_{i}^{n} (\alpha_i+\alpha_i^*) \tag{4}$$

with constraints

$$\alpha_i, \alpha_i^* \in [0, C].$$

Solving the above equation with the constraints determines the optimal Lagrange multipliers, α_i, α_i^*, the estimator of the regression function given the input vector $\boldsymbol{x}$ are obtained as follows.

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} (\alpha_i-\alpha_i^*)K(\boldsymbol{x}_i, \boldsymbol{x}). \tag{5}$$

In the nonlinear case, $\boldsymbol{w}$ is no longer explicitly given. However, it is uniquely defined in the weak sense by the dot products. Here the linear regression model can be regarded as the special case of the nonlinear regression model by using identity feature mapping function, that is, $\phi(\boldsymbol{x}) = \boldsymbol{x}$ which implies the linear kernel such that $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \boldsymbol{x}_1'\boldsymbol{x}_2$.

3. e-SVR using IRWLS

In this section we propose e-SVR using IRWLS which is an IRWLS procedure to solve the QP problem of e-SVR with a modified loss function which is differentiable at $\pm e$. The modified loss function $\rho_{e,\delta}(\cdot)$ is attained by providing the differentiability at $\pm e$ by differing from the original e-insensitive loss function $\rho_e(\cdot)$ in the interval $(-e-\delta, e+\delta)$,

$$\begin{aligned}\rho_{e,\delta}(r) = (-r-e)\ \ I(r < e-\delta) + \frac{\delta r^2}{(e+\delta)^2} I(-e-\delta \leqq r \leqq e+\delta) \\ + (r-e)\ \ I(r > e+\delta),\end{aligned} \tag{6}$$

where $I(\cdot)$ is an indicative function.

The representation theorem(Kimeldorf and Wahba, 1971) guarantees the minimizer of the optimization problem (1) to be $\hat{f}(\boldsymbol{x}) = K\beta$ for some vector $\beta = \alpha - \alpha^*$. Now the problem (1) becomes obtaining β to minimize

$$L(\beta) = \frac{1}{2}\beta' K\beta + \frac{C}{2}\sum_{i=1}^{n} \rho_{e,\delta}(y_i - K_i\beta). \tag{7}$$

where K_i is the i-th row of K. Taking partial derivatives of (7) with regard to β leads to the optimal values of β to be the solution to

$$\mathbf{0} = K\beta - CKW\boldsymbol{y} + CKWK\beta\ . \tag{8}$$

Here W is a diagonal matrix with the i-th diagonal element w_{ii} obtained from the derivative of the modified loss function as

$$\begin{aligned}w_{ii} = \frac{-1}{r_i} I(r_i < -e-\delta) + \frac{2\delta}{(e+\delta)^2} I(-e-\delta \leqq r_i \leqq -e-\delta) \\ + \frac{1}{r_i} I(r_i > e+\delta)\end{aligned} \tag{9}$$

where $r_i = y_i - K_i\beta$.

The solution to (8) cannot be obtained in a single step since W contains β. Thus we need to apply IRWLS procedure which starts with initialized values of β as follows:

(a) Calculate W with β.

(b) Calculate β from $\beta = (KWK + K/C)^{-1}KW\boldsymbol{y}$.

(c) Iterate steps until convergence.

The functional structures of e-SVR is characterized by hyperparameters - the regularization parameter C and the kernel parameters. The cross validation(CV) technique used in SVR with the quadratic loss function cannot be used in e-SVR since the loss function used in e-SVR is not differentiable as the quadratic loss function. To select the parameters of e-SVR using IRWLS we consider the cross validation(CV) function as follows:

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \rho_{e,\delta}(y_i - \hat{f}_\lambda^{(-i)}(\boldsymbol{x}_i)),$$

where λ is the set of hyperparameters and $\hat{f}_\lambda^{(-i)}(\boldsymbol{x}_i)$ is the regression function estimated without i-th observation. Since for each candidates of hyperparameters, $\hat{f}_\lambda^{(-i)}(\boldsymbol{x}_i)$ for $i = 1, \cdots, n$, should be evaluated, selecting parameters using CV function is computationally formidable. Using derivation of ACV function from CV function by Nychka et al.(1995), we have GACV function as follows

$$GACV(\lambda) = \frac{\sum_{i=1}^{n} \rho_{e,\delta}(y_i - \hat{f}_\lambda(\boldsymbol{x}_i))}{n - tr(H)}, \tag{10}$$

where H is the hat matrix such that $\hat{f}_\lambda(\boldsymbol{x}) = H\boldsymbol{y}$ with the (i,j)-th element $h_{ij} = \partial\hat{f}(\boldsymbol{x}_i)/\partial y_j$. GACV function cannot be applied to e-SVR using QP since H is not computable. But for e-SVR using IRWLS, hyperparameters can be selected by applying (10), where H is obtained easily as ,

$$H = K(KWK + K/C)^{-1}KW. \tag{11}$$

4. Numerical Studies

We illustrate the performance of the regression estimation using e-SVR using IRWLS through the simulated example on the nonlinear regression cases.
101 data sets are generated to present the prediction performance of the proposed procedure - one for training and 100 for testing. Each data set consists of 100 x's and 100 y's. Here x's are equally spaced ranging from 0 to 1 and y's are generated from a normal distribution $N(1+sin(2\pi x), 0.01)$. The regression function a given $\boldsymbol{x}$ can be modelled as $\hat{f}(\boldsymbol{x}) = \boldsymbol{w}'\phi(\boldsymbol{x})$ where $\boldsymbol{x} = (1, x)'$. True regression function is given as $f(\boldsymbol{x}) = 1 + sin(2\pi x)$ We set δ in the modified loss function (8) to 0.001 and e to 0.1(Kwok and Tsang(2003) suggested the optimal value of $e \approx$ standard deviation of regression error). We stop iterations when mean squared difference of two successive regression function estimates is less than 0.000001. The radial basis kernel function is utilized in this example, which is

$$K(x_1, x_2) = exp(-\frac{1}{\sigma^2}(x_1 - x_2)^2).$$

We select (C, σ^2) as (100, 1.0) using GACV function (10). To illustrate the prediction performance of e-SVR using IRWLS, we compare it with e-SVR using QP via 100 test data sets, where the predicted mean squared error(PMSE) is used as prediction performance measure defined by

$$PMSE = \frac{1}{n_t}\sum_{i=1}^{n_t}(\hat{f}(\boldsymbol{x}_{ti}) - f(\boldsymbol{x}_{ti}))^2.$$

The averages of 100 PMSEs from e-SVR using IRWLS and QP are obtained as 0.001913 and 0.002376, respectively, which implies that both procedures have almost same rediction performance. Figure 1 shows the true regression function (solid line) for one of 100 test data sets and the regression function predictors(dotted line) obtained by e-SVR using IRWLS and QP, respectively. We can see that the proposed procedure works reasonably well for the nonlinear regression.

Figure 1. True Regression Function and Regression Function Predictors obtained by e-SVR using IRWLS(Left) and e-SVR using QP(Right). The scatter is 100 artificial data points (x_t, y_t) of test data set with x_t's are equally spaced ranging from 0 to 1, and y_t's generated from a normal distribution $N(1+sin(2\pi x), 0.01)$. True regression function(solid line) and the regression function predictor(dotted line) are superimposed on the scatter plots.

With simulated data sets, CPU-time of the proposed procedure are compared with that of e-SVR using QP computed by the built-in function of MATLAB. Here (C, σ^2) are fixed as (100, 1.0). Table1 shows CPU-time in seconds of both procedures (run MATLAB 6.5 over Pentium IV at 2.0GHz) for e-SVR on a data set with different sample sizes. From table1 we can see that e-SVR using IRWLS is much faster than e-SVR using QP, which implies that the proposed procedure is appropriate procedure for the large training data sets.

Table 1. CPU time in seconds for training e_SVR using IRWLS and QP

sample size	IRWLS	QP	sample size	IRWLS	QP
100	1.8	17.3	300	37.9	781.7
200	11.6	215.8	400	84.4	2000.0

4. Conclusions

In this paper, we dealt with estimating the regression function by e-SVR using IRWLS and obtained GACV function for the proposed procedure. Through the examples we showed that the proposed procedure derives the satisfying solutions. We also found that e-SVR using IRWLS is much faster than e-SVR using QP, which implies that the proposed procedure is appropriate for the large training data sets.

References

1. Kimeldorf, G. S. and Wahba, G. (1971). Some Results on Tchebycheffian Spline Functions. *Journal of Mathematical Analysis and its Applications,* (33), 82-95.
2. Kwok, J. T. and Tsang, I .W. (2003). Linear dependency between epsilon and the input noise in epsilon-support vector regression. *IEEE Transactions on Neural Networks*, 14, 3, 544-553.
3. 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.
4. Nychka, D., Gray, G., Haaland, P., Martin, D., O'Connell, M. (1995). A Nonparametric Approach Syringe Grading for Quality Improvement. *Journal of American Statistical Association*, 432, 1171-1178.
5. Perez-Cruz, F., Navia-Vazquez, A., Alarcon-Diana, P. L. ,and Artes-Rodriguez, A. (2000). An IRWLSprocedure for SVR. *In Proceedings of European Association for Signal Processing*, EUSIPO 2000, Tampere, Finland.
6. Platt, J. (1998). Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Microsoft Research Technical Report MSR-TR-98-14.
7. Smola, A. and Scholkopf, B. (1998). On a Kernel-Based Method for Pattern Recognition, Regression, Approximation and Operator Inversion. *Algorithmica*, 22, 211-231.
8. Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. *Springer*, New York.
9. Vapnik, V. N. (1998). Statistical Learning Theory. *John Wiley*, New York.
10. Wang, L.(Ed.) (2005). Support Vector Machines: Theory and Application. *Springer*, Berlin Heidelberg New York.

[received date : Sep. 2005, accepted date : Oct. 2005]