Journal of Korean
Data & Information Science Society
2007, Vol. 18, No. 3, pp. 735 ~ 744

Sparse Kernel Regression using IRWLS Procedure

Hyejung Park[1]

Support vector machine(SVM) is capable of providing a more complete description of the linear and nonlinear relationships among random variables. In this paper we propose a sparse kernel regression(SKR) to overcome a weak point of SVM, which is, the steep growth of the number of support vectors with increasing the number of training data. The iterative reweighted least squares(IRWLS) procedure is used to solve the optimal problem of SKR with a Laplacian prior. Furthermore, the generalized cross validation(GCV) function is introduced to select the hyper-parameters which affect the performance of SKR. Experimental results are then presented which illustrate the performance of the proposed procedure.

Keywords : Generalized Cross Validation Function, Iterative Reweighted Least Squares Procedure, Kernel Function, Laplacian Prior, Sparsity, Support Vector, Support Vector Regression

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 the 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 can be found in Vapnik(1995,1998), Smola and Schö lkopf(1998), and Wang(2005). Training an SVM requires the solution to a quadratic programming(QP) optimization problem. But QP problem presents some inherent limitations which result in computational difficulty especially for the large data

[1] Faculty of Mobile Contents, Haany University of Daegu, Kyungbuk, 712-715, Korea
E-mail : hyjpark@dhu.ac.kr

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 SVM by transforming the Lagrangian function into sum of quadratic terms by defining associated weights of predicted errors.

Sparsity is known as an important feature of kernel regression models, which provides the efficiency on predicting the regression function. SVM does not provide extreme sparsity and the number of support vectors depends on the number of training data. Tipping(2001) proposed a Bayesian approach referred to as the relevance vector machine(RVM) providing more sparsity. However RVM has computational problems since there are no closed-form solutions for maximizing the marginal likelihood.

In this paper we propose a SKR using IRWLS procedure to obtain simultaneously the accuracy and the sparsity. Also the proposed SKR enables to select appropriate hyper-parameters easily from the generalized cross validation(GCV) function, which is used to select hyper-parameters for the achievement of high generalization performance. The rest of this paper is organized as follows. In Section 2 we give brief reviews of SVM and RVM for regression. In Section 3 we propose a SKR using IRWLS procedure and present the model selection method using GCV function. In Section 4 we perform the numerical studies through examples. In Section 5 we give the conclusions.

2. Kernel Regressions

2.1 Support Vector Machine

Let the training data set denoted by $(x_{i,} y_i)_{i=1}^n$, with each input $x_i \in R^d$ including a constant 1 and the response $y_i \in R$, where the output variable y_i is related to the input vector 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 function in input space, $\phi(x_i)'\phi(x_j) = K(x_i, x_j)$ (Mercer, 1909). We consider the nonlinear regression case, in which the regression function of the response given x, $\mu(x)$, can be regarded as a nonlinear function of input vector 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,

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

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

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

subject to

$$y_i - w'\phi(x_i) \leqq e + \xi_i \tag{3}$$
$$w'\phi(x_i) - y_i \leqq e + \xi_i^*, \quad e, \xi_i, \xi_i^* \geqq 0$$

where C is a regularization parameter penalizing the training errors.
We construct a Lagrange function from (2) and (3) as follows:

$$L = \frac{1}{2}w'w + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}\alpha_i(e + \xi_i - y_i + w'\phi(x_i)) \tag{4}$$
$$- \sum_{i=1}^{n}\alpha_i^*(e + \xi_i^* + y_i - w'\phi(x_i)) - \sum_{i=1}^{n}(\eta_i\xi_i + \eta_i^*\xi_i^*).$$

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 (4) with regard to the primal variables (w, ξ_i, ξ_i^*) and plugging them into equation (4), we have the optimization problem below.

$$\max -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(x_i, x_j) + \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)y_i - e\sum_{i}(\alpha_i + \alpha_i^*)$$

with constraints

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

where the data points corresponding to positive values of α_i or α_i^* are called support vectors. 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 x are obtained as follows.

$$\hat{\mu}(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x_i, x).$$

In the nonlinear case, 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, $\phi(x) = x$ which implies the linear kernel such that $K(x_1, x_2) = x_1'x_2$.

2.2 Relevance Vector Machine.

Given the training data set $(x_i, y_i)_{i=1}^{n}$, with each input $x_i \in R^d$ and the response $y_i \in R$, the distribution of y given x is assumed to follow a normal distribution, $N(\mu(x), \sigma^2)$, where the mean is modelled by $\mu(x)$ as defined in (1) for SVR. Then the likelihood of training data can be expressed as

$$p(y|\,\alpha,\sigma^2) = (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\,\|\,\boldsymbol{y}-\Phi\,\alpha\,\|^2\right),$$

where $\alpha = (\alpha_0,\cdots,\alpha_n)'$ and Φ is a $n\times(n+1)$ design matrix consisted of $\mathbf{1}_n$ and the kernel function K with $K_{ij}=K(\boldsymbol{x}_i,\boldsymbol{x}_j)$. The normal prior is imposed over the weight vector α,

$$p(\alpha) = \prod_{i=0}^{n} N(0,1/a_i),$$

with $n+1$ hyper-parameters a_i's.

Then posterior distribution over the weights is obtained by Bayes' rule as follow,

$$p(\boldsymbol{\alpha}|\boldsymbol{y},\boldsymbol{a},\sigma^2)=(2\pi\sigma^2)^{-(n+1)/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\mu})'|\Sigma|^{-1}(\boldsymbol{\alpha}-\boldsymbol{\mu})\right),$$

where $\Sigma=(\Phi' B\Phi+diag(\boldsymbol{a}))^{-1}$, $\boldsymbol{\mu}=\Sigma\Phi' B\boldsymbol{y}$, and $B=\mathrm{I_n}/\sigma^2$.

By integrating out the weights $\boldsymbol{\alpha}$, the marginal likelihood for hyper-parameters is obtained as

$$p(\boldsymbol{y}|\boldsymbol{a},\sigma^2)=(2\pi\sigma^2)^{-n/2}|B^{-1}+\Phi A^{-1}\Phi'|^{-1/2}\exp\left(-\frac{1}{2}\boldsymbol{y}'(B^{-1}+\Phi A^{-1}\Phi')^{-1}\boldsymbol{y}\right).$$

Relevance vector learning becomes to find the hyper-parameter posterior mode by maximizing $p(\boldsymbol{a},\sigma^2\,|\,\boldsymbol{y})\propto p(\boldsymbol{y}|\boldsymbol{a},\sigma^2)p(\boldsymbol{a})p(\sigma^2)$ with respect to $\boldsymbol{a}$ and σ^2. Since the maximizing values of $\boldsymbol{a}$ and σ^2 cannot be obtained in a closed form, the iterative re-estimation procedure is used(MacKay, 1992).

With the optimal values of $\boldsymbol{a}$, the estimator of the regression function given the input vector $\boldsymbol{x}$ are obtained as follows.

$$\hat{\mu}(\boldsymbol{x})=\sum_{i=1}^{n} a_i\,\Phi(\boldsymbol{x}_i,\boldsymbol{x})\,.$$

3. Sparse Kernel Regression using IRWLS procedure

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 $y_i \in R$. For this data set, we can consider the regression model

$$y_i = \mu(\boldsymbol{x}_i)+\epsilon_i,\ \ i=1,2,\cdots,n,$$

where ϵ_i is assumed to be independently normally distributed with mean 0 and variance σ^2 and μ is the regression function to be estimated.

The negative log likelihood of the given data set can be expressed as(constant terms are omitted)

$$\ell\,(\mu\,|\,\boldsymbol{x})=\frac{1}{2n}\sum_{i=1}^{n}(y_i-\mu(\boldsymbol{x}_i))^2.$$

The regression function is estimated by a linear model, $\mu(\boldsymbol{x}_i)=\omega'\phi(\boldsymbol{x}_i)$ conducted in a high dimensional feature space, which can be rewritten as $\mu(\boldsymbol{x}_i)=K_{i.}\boldsymbol{\alpha}$, where $K_{i.}$ is the i -th row of K and α is the vector of n weights to be estimated. Then the maximum likelihood estimates of α are obtained by minimizing the negative log-likelihood function,

$$\ell\,(\alpha)=\frac{1}{2n}\sum_{i=1}^{n}(y_i-K_{i.}\alpha)^2.$$

The maximum likelihood estimates of α generally lead severe overfitting, we are encouraged to use a prior over α . Then the penalized maximum likelihood estimates(the maximum a posteriori estimates) of α are obtained by minimizing the objective function,

$$L(\alpha)=\ \ell\,(\alpha)+\log p(\alpha),$$

where $p(\alpha)$ is some prior over α.

To have the sparsity on estimation of α , we use a Laplacian prior(Williams, 1995),

$$p(\alpha)\propto\exp(-\lambda\,\|\alpha\|_1),$$

where $\|\alpha\|_1=\sum_i|\alpha_i|$ denotes ℓ_1 norm and λ is a nonnegative constant.

The objective function can be rewritten as

$$L(\alpha)=\ \ell\,(\alpha)+\ \lambda\|\alpha\|_1\,.$$

Here λ controls the tradeoff between the goodness-of-fit on the data and $\|\alpha\|_1$.

The objective function $L(\alpha)$ is not differentiable with respect to α , we need a modification of $L(\alpha)$ for IRWLS procedure.

We define an objective function given α^* as

$$L(\alpha\,|\,\alpha^*)=\ \ell\,(\alpha)+\frac{\lambda}{2}\sum_{i=1}^{n}\left(\frac{\alpha_i^2}{|\alpha_i^*|}+|\alpha_i^*|\right)\ ,$$

then $L(\alpha\,|\,\alpha^*)\geq L(\alpha)$ with equality if and only if $\alpha=\alpha^*$ (Krishnapuram et al., 2005) and $L(\alpha\,|\,\alpha^*)$ is differentiable with respect to α.

At t-th iteration of IRWLS procedure, we have

$$L(\alpha\,|\,\alpha^{(t)})=\ \ell\,(\alpha)+\frac{\lambda}{2}\sum_{i=1}^{n}\left(\frac{\alpha_i^2}{|\alpha_i^{(t)}|}+|\alpha_i^{(t)}|\right)\ .$$

Then $\alpha^{(t+1)}$ is obtained by minimizing $L(\alpha\,|\,\alpha^{(t)})$ with respect to α as

$$\alpha^{(t+1)}=\ (KK+\lambda\,W(\alpha^{(t)}))^{-1}Ky\ ,$$

where $W(\alpha^{(t)})$ is the diagonal matrix consisted of $1/|\alpha_i^{(t)}|\,,\ i=1,\cdots,n$.

During iteration, we find that some α_i's tend to zero keeping the value of objective function $L(\alpha)$ decreasing. This motivates that we can find sparse estimates of α which provides decreasing value of the objective function $L(\alpha)$ at the same time.

```
Algorithm of SKR using IRWLS Procedure:
1. Set $v = (1:n)'$ and $\boldsymbol{\alpha}(v)^{(0)}$ .
2. Find solution $\boldsymbol{\alpha}(v)^{(t+1)}$ which minimizes $L(\boldsymbol{\alpha}(v) \mid \boldsymbol{\alpha}(v)^{(t)})$ .
3. Set $\alpha_i^{(t+1)} = 0$ which is very close to zero.
Find $v = \{i \mid \alpha_i^{(t+1)} \neq 0\}$, which is, a vector of subset of $\{1, 2, \cdots, n\}$ satisfying $\alpha_i^{(t+1)} \neq 0$ .
4. iterate 2-4 until $\mid L(\boldsymbol{\alpha}(v)^{(t+1)}) - L(\boldsymbol{\alpha}(v)^{(t)}) \mid <$ Tol.
```

The functional structures of SKR is characterized by hyper-parameters, the regularization parameter C and the kernel parameters. To select the hyper-parameters of SKR using IRWLS we define the cross validation(CV) function as follows:

$$CV(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_{\boldsymbol{\lambda}}{}^{(-i)}(\boldsymbol{x}_i))^2 ,$$

where $\boldsymbol{\lambda}$ is the set of hyper-parameters and $\hat{\mu}_{\boldsymbol{\lambda}}{}^{(-i)}(\boldsymbol{x}_i)$ is the regression function estimated without i-th observation. Since for each candidate of hyper-parameters, $\hat{\mu}_{\boldsymbol{\lambda}}{}^{(-i)}(\boldsymbol{x}_i)$ for $i = 1, \cdots, n$, should be evaluated, selecting parameters using CV function is computationally burdensome. Using derivation of the generalized approximate cross validation function from the CV function by Yuan(2006), we have the generalized cross validation(GCV) function as follows

$$GCV(\boldsymbol{\lambda}) = \frac{n\sum_{i=1}^{n}(y_i - \hat{\mu}_{\boldsymbol{\lambda}}(\boldsymbol{x}_i))^2}{(n - tr(H))^2}, \tag{5}$$

where $H = K(:,v)(K(:,v)'K(:,v) + \lambda W)^{-1}K(:,v)', v = \{i : \alpha_i \neq 0\}$, is the hat matrix such that $\hat{\mu}_{\boldsymbol{\lambda}}(\boldsymbol{x}) = H\boldsymbol{y}$ with the (i,j)-th element $h_{ij} = \partial\hat{\mu}(\boldsymbol{x}_i)/\partial y_j$. GCV function cannot be applied to SVR using QP since H is not computable. But for SKR using IRWLS, hyper-parameters can be easily selected by applying GCV function.

4. Numerical Studies

We illustrate the performance of the regression estimation using SKR using IRWLS through the simulated data and the real data on the nonlinear regression cases.

200 data sets are generated to present the prediction performance of the proposed procedure - 100 for training and 100 for testing. Each data set consists of 100 x's and 100 y's. Here x's are randomly generated from $U(0,1)$ and y's are generated from a normal distribution $N(1 + x + \sin(2\pi x), 1)$. Figure 1(Left) shows the true regression functions imposed on the scatter plots of one of 100 training

data sets. The regression function a given x can be modelled as $\mu(x) = w'\phi(X)$ where $X = (1,x)'$. The radial basis kernel function is utilized in this example, which is

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

For SKR using IRWLS, (C,σ^2) is selected from GCV function (5). To illustrate the prediction performance of SKR using IRWLS, we compare it with SVM and RVM. For SVM, e is obtained from $3s\sqrt{\frac{\log(n)}{n}} = 0.839$, where s is the standard deviation of y (Cherkassky and Ma, 2004). 10 fold cross validation is used in SVM and RVM for the selection of (C,σ^2) and σ^2, respectively.
The predicted mean squared error(PMSE) is used as the prediction performance measure defined by

$$PMSE = \frac{1}{n_t}\sum_{i=1}^{n_t}(\mu(x_{ti}) - \hat{\mu}(x_{ti}))^2.$$

The averages of 100 PMSEs and the averages of numbers of retained kernel functions by SVM, RVM, and SKR using IRWLS are obtained as (0.0757, 0.1257, 0.0551) and (54.32, 11.46, 40.07), respectively. Figure 2(Left) shows boxplots of PMSEs obtained by SVM, RVM, and SKR using IRWLS, respectively. We can see that SKR using IRWLS, provides better result than other two regressions in this example. Figure 2(Right) shows boxplots of numbers of retained kernel functions. We can see that SKR using IRWLS provides more sparsity than SVM.

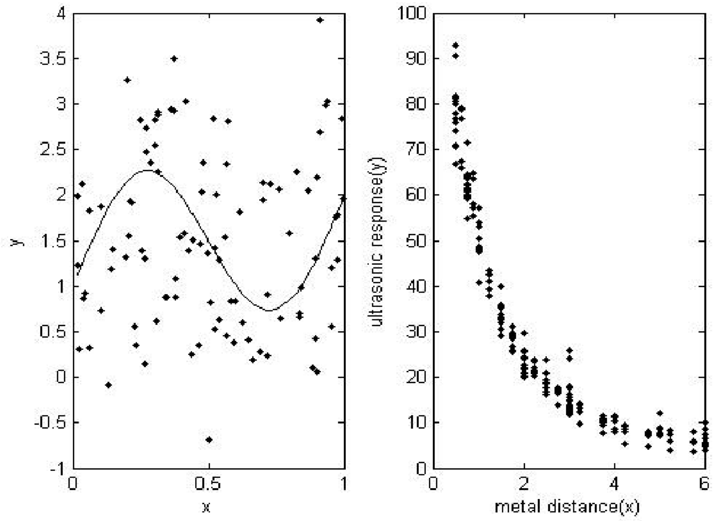

Figure 1. True regression functions imposed on the scatter plots of one of 100 training data sets(Left) and the scatter plots of Ultrasonic data(Right).

The ultrasonic data set available from http://www.itl.nist.gov/div898/strd/nls, where the ultrasonic response and the metal distance are known to be nonlinearly related, is used to illustrate the prediction performance of SKR using IRWLS. Figure 1(Right) shows the scatter plots of 214 data points. We randomly divided the data into training data(142 data points) and test data(72 data points) 100 times. The radial basis kernel function is utilized in this example. For SKR using IRWLS, (C, σ^2) is selected from GCV function (5). To illustrate the prediction performance of SKR using IRWLS, we compare it with SVM and RVM. For SVM, e is obtained as 0.5604(Cherkassky and Ma, 2004). 10 fold cross validation is used in SVM and RVM for the selection of (C, σ^2) and σ^2, respectively. Here the predicted mean squared error(PMSE) is defined by

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

The averages of 100 PMSEs and the averages of numbers of retained kernel functions by SVM, RVM, and SKR using IRWLS are obtained as (11.0885, 11.8355, 11.5905) and (114.83, 38.98, 81.78), respectively. Figure 3(Left) shows boxplots of PMSEs obtained by SVM, RVM, and SKR using IRWLS, respectively. We can see that SVM provides slightly better result than other two regressions in this example. Figure 3(Right) shows boxplots of numbers of retained kernel functions. We can see that SKR using IRWLS provides more sparsity than SVM.

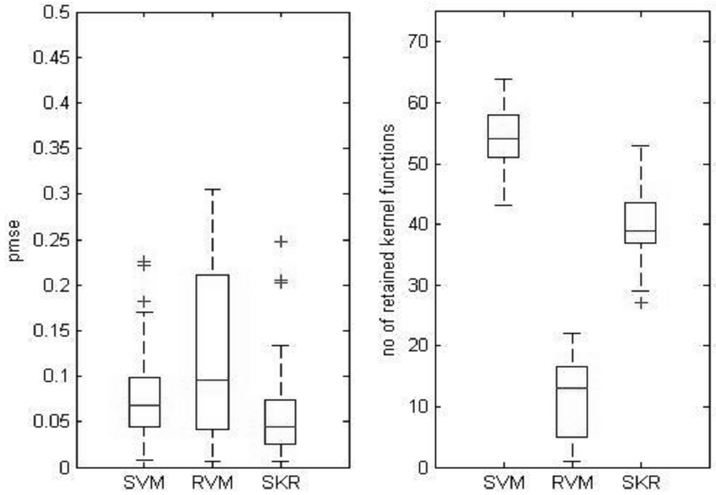

Figure 2. Predicted mean squared errors(Left) and numbers of retained kernel functions(Right) by SVM, RVM, and SKR using IRWLS, respectively, for simulated data.

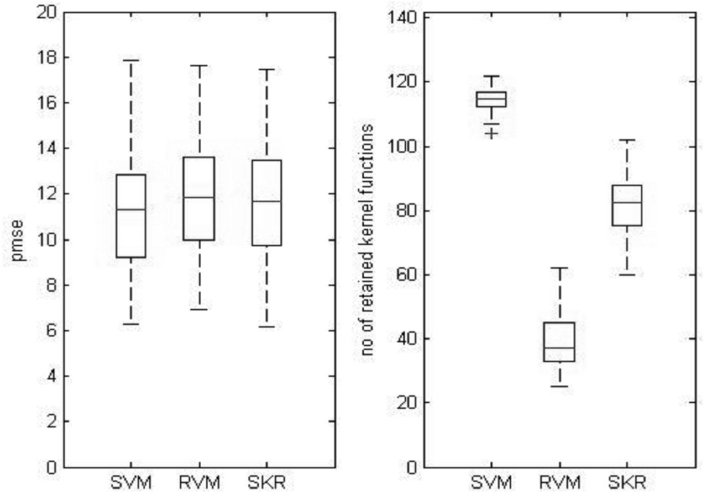

Figure 3. Predicted mean squared errors(Left) and numbers of retained kernel functions(Right) by SVM, RVM, and SKR using IRWLS, respectively, for Ultrasonic data.

5. Conclusions

In this paper, we dealt with estimating the regression function by SKR using IRWLS and obtained GCV function for the proposed procedure. Through the example we showed that the proposed procedure derives the satisfying results. We also found that SKR using IRWLS has an advantage other than SVM and RVM, that is, it provides an easy model selection method using GCV function.

References

1. Cherkassky, V. and Ma, Y. (2004). Practical Selection of SVM Parameters and Noise Estimation for SVM Regression. *Neural Networks*, 17, 1, 113-126.
2. Krishnapuram, B., Carlin, L., Figueiredo, M. A. T., and Hartermink, A. J. (2005). Sparse Multinomial Logistic Regression: Fast Algorithms and Generalization Bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, 6, 957-968.
3. Mackay, D. J. C. (1992). Bayesian Interpolation. *Neural Computation*, 4(3), 415-447.
4. 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.

5. Perez-Cruz, F., Navia-Vazquez, A., Alarcon-Diana, P. L., and Artes-Rodriguez, A. (2000). An IRWLS procedure 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. Tipping, M. E. (2001). Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1, 211-244.
9. Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. *Springer*, New York.
10. Vapnik, V. N. (1998). Statistical Learning Theory. *John Wiley*, New York.
11. Wang, L.(Ed.) (2005). Support Vector Machines: Theory and Application. *Springer*, Berlin Heidelberg New York.
12. Williams, P. M. (1995). Bayesian Regularization and Pruning Using a Laplace Prior. *Neural Computation*, 7, 117-143.
13. Yuan, M. (2006). GACV for Quantile Smoothing Splines. *Computational Statistics & Data Analysis*, 50(3), 813-829.

[received date : Mar. 2007, accepted date : June 2007]