Asian Journal of Information Technology 4 (11): 978-980, 2005

PKSVR: A Novel Prior Knowledge-based Support Vector Regression

[1,2]Wensen An and [2]Yanguang Sun
[1]Department of automation, University of Science and Technology of China, HeFei 230027, China
[2]Automation Research and Design Institute of Metallurgical Industry, Beijing 100071, China

Abstract: Standard formulation of support vector regression is first introduced, and then a modification of classical support vector regression is discussed detailedly by incorporating prior knowledge in the form of inequalities to be satisfied by the desired regression function as a means of compensating for a shortage of training data. Experimental result shows that the proposed method can significantly improve the performance of support vector regression.

Key words: Prior knowledge, support vector regression, support vector machines

INTRODUCTION

Support Vector Machines (SVMs) have been successfully applied in many real-world applications [1,3]. More recently, prior knowledge has been incorporated into SVMs classifiers [4,5] both to improve the classification task and to handle problems where conventional data may be few or not available. Schölkopf [5] showed that the prior knowledge can be incorporated with the appropriate kernel function, and Fung [4] showed prior knowledge in the form of multiple polyhedral sets can be used with a reformulation of SVMs. However, little work [6,7] has been done to incorporate prior knowledge into Support Vector Regression (SVR) as has been done for SVMs classifier though SVMs have also been extensively used for regression [8-10].

In its standard form, SVR does not allow for the direct incorporation of prior knowledge. In this study we describe a new modification of standard SVR. The basic idea of our approach is to introduce prior knowledge in the form of inequalities to be satisfied by the desired regression function, and incorporate them into the standard SVR algorithm.

STANDARD FORMULATION OF SVR

Let $D=\{(x_k,y_k)\,|\,k=1,\cdots,m\}$ $x_k\in R^n$ $y_k\in R$ be a given training data set of input x_k and associated targets y_k. The goal of the regression problem is to fit a flat function f(x) which approximates the relation inherited between the data set points and it can be used later on to infer the output y for a new input data point x.

Suppose the function f(x) is expressed as:

$$f(x)=<\omega,\phi(x)>+b \tag{1}$$

where $\leq,\geq$ is dot product of vector, $b\in R$ is a bias term, and $\phi: R^n\rightarrow F$ is a nonlinear map which mapping the input x into a high-dimensional feature space F.

According to SRM principle, that function f(x) is flat in the case of Eq. 1 means that one seeks the minimization of the following expression:

$$\frac{1}{2}\|\omega\|^2+C\sum_{k=1}^{m}L(f(x_k),y_k) \tag{2}$$

where L(.) is a loss function, C is a constant.

Many forms for the loss function can be found in the literature: e.g. linear, quadratic and huber loss function, etc. In this paper, Vapnik's loss function is used, which is known as ε- insensitive loss function and defined as:

$$L(y,f(x))=\begin{cases}0 & |f(x)-y|<\varepsilon\\ |y-f(x)|-\varepsilon & \text{otherwise}\end{cases} \tag{3}$$

Then the regression problem can be written as a convex optimization problem as follows:

$$\begin{aligned}&\min\ \frac{1}{2}\|\omega\|^2+C\sum_{k=1}^{m}(\xi_k+\xi_k^*)\\ &\text{s.t.}\begin{cases}y_k-<\omega,\phi(x_k)>-b\leq\varepsilon+\xi_k\\ <\omega,\phi(x_k)>+b-y_k\leq\varepsilon+\xi_k^*\\ \xi_k,\xi_k^*\geq 0\end{cases}\end{aligned} \tag{4}$$

$\varepsilon>0$ is a predefined constant which controls the noise tolerance, the constant $C>0$ determines the trade-off between the flatness of f and the amount of tolerable deviations, which is larger than ε.

Through introducing a Lagrange function, the optimization problem (4) can be solved in their dual formulation, which is expressed as follows:

$$\begin{aligned}&\max\ -\frac{1}{2}\sum_{k,l=1}^{m}(\alpha_k-\alpha_k^*)(\alpha_l-\alpha_l^*)<\phi(x_k),\phi(x_l)>\\ &-\varepsilon\sum_{k=1}^{m}(\alpha_k+\alpha_k^*)+\sum_{k=1}^{m}y_k(\alpha_k-\alpha_k^*)\ \text{s.t.}\begin{cases}\sum_{k=1}^{m}(\alpha_k-\alpha_k^*)=0\\ \alpha_k,\alpha_k^*\in[0,C]\end{cases}\end{aligned} \tag{5}$$

Corresponding Auhtor: Wensen An,Department of automation, University of Science and Technology of China, HeFei 230027, China

The optimal value of α_k, α_k^* can be obtained by solving the dual problem (5), accordingly, the ω and f(x) can be described by:

$$\begin{aligned} \omega &= \sum_{k=1}^{m} (\alpha_k - \alpha_k^*)\phi(x_k) \\ f(x) &= \sum_{k=1}^{m} (\alpha_k - \alpha_k^*) < \phi(x_k), \phi(x) > + b \end{aligned} \tag{6}$$

and the value of b can be computed according to the Karush-Kuhn-Tucker (KKT) conditions.

By introducing kernel function instead of nonlinear mapping ϕ due to unnecessary to know ϕ explicitly, i.e. $K(x,x') = < \phi(x), \phi(x') >$, f(x) is rewritten as follows:

$$f(x) = \sum_{k=1}^{m} (\alpha_k - \alpha_k^*) K(x_k, x) + b \tag{7}$$

As above stated, the prior knowledge of problem is not utilized in standard SVR algorithm. Next we will discuss how prior knowledge can be incorporated into standard SVR algorithm.

PRIOR KNOWLEDGE-BASED SVR

In practical applications, we usually can obtain some prior knowledge that depict problem to be solved. Now we describe Prior Knowledge-based Support Vector Regression (PKSVR) formulation.

Suppose prior knowledge can be described in the following form of inequalities to be satisfied by the desired regression function:

$$< p_i, y(x) > \le q_i \quad i = 1, 2, \cdots N \tag{8}$$

where $p_i = [p_{i1}, p_{i2}, \cdots, p_{im}]^T, q_i \in R$, $Y(X) = [y(x_1), y(x_2), \cdots, y(x_m)]^T$

$$y(x_i) = < \omega, \phi(x_i) > + b$$

The value of p_i and q_i can be decided by the prior knowledge of special problem. Obviously it's easy to extend (8) to the following circumstances:

(I) $u \le y(x_i) \le v$, u and v is the lower and upper bound, respectively;(ii) $y(x_i) - y(x_j) \le 0$

We introduce prior knowledge with coupling (8) to the optimization problem (4), and then a new optimization problem can be obtained as follows:

$$\begin{aligned} &\min \ \frac{1}{2}\|\omega\|^2 + C\sum_{k=1}^{m} (\xi_k + \xi_k^*) \\ &\text{s.t.} \begin{cases} y_k - < \omega, \phi(x_k) > - b \le \varepsilon + \xi_k \\ < \omega, \phi(x_k) > + b - y_k \le \varepsilon + \xi_k^* \\ \xi_k, \xi_k^* \ge 0 \\ < p_i, y(x) > \le q_i \ \ i = 1, 2, \cdots N \end{cases} k = 1, 2, \cdots m \end{aligned} \tag{9}$$

The optimization problem (9) re transformed to the Lagrange function as follows:

$$\begin{aligned} L(\omega, b, \xi, \xi^*, \alpha, \alpha^*, \beta, \beta^*, \gamma) &= \frac{1}{2}\|\omega\|^2 + C\sum_{k=1}^{m} (\xi_k + \xi_k^*) - \\ &\sum_{k=1}^{m} \alpha_k [\varepsilon + \xi_k - y_k + < \omega, \phi(x_k) > + b] - \sum_{k=1}^{m} \alpha_k^* [\varepsilon + \xi_k^* + y_k - \\ &< \omega, \phi(x_k) > - b] - \sum_{k=1}^{m} (\beta_k \xi_k + \beta_k^* \xi_k^*) - \sum_{k=1}^{m} \gamma_k [q_k - < p_k, y(x) >] \end{aligned} \tag{10}$$

It follows form the saddle point condition that the partial derivatives of L with respect to the variables (ω, b, ξ, ξ^*) have to vanish for optimality:

$$\begin{aligned} \frac{\partial L}{\partial \omega} &= \omega - \sum_{k=1}^{m} (\alpha_k - \alpha_k^* - \sum_{i=1}^{N} \gamma_i p_{ik}) \phi(x_k) = 0 \\ \frac{\partial L}{\partial b} &= -\sum_{k=1}^{m} \alpha_k + \sum_{k=1}^{m} \alpha_k^* + \sum_{k=1}^{m} \sum_{i=1}^{N} \gamma_i p_{ik} = 0 \\ \frac{\partial L}{\partial \xi_k} &= C - \alpha_k - \beta_k = 0 \\ \frac{\partial L}{\partial \xi_k^*} &= C - \alpha_k^* - \beta_k^* = 0 \end{aligned} \tag{11}$$

Substituting (11) into (10) yields the dual optimization problem:

$$\max \ -\frac{1}{2}\sum_{k,l=1}^{m} \rho_k \rho_l K(x_k, x_l) + \sum_{i=1}^{N} \gamma_i q_i - \sum_{k=1}^{m} \varepsilon(\alpha_k + \alpha_k^*) + \sum_{k=1}^{m} y_k (\alpha_k - \alpha_k^*)$$

$$\text{s.t.} \begin{cases} \rho_k = \alpha_k - \alpha_k^* - \sum_{i=1}^{N} \gamma_i p_{ik} \\ \sum_{k=1}^{m} \rho_k = 0 \\ \alpha_k, \alpha_k^* \in [0, C] \\ \gamma_i \ge 0 \end{cases} \tag{12}$$

In the same way as standard SVR algorithm, the optimal value of α_k, α_k^* and γ_i, can be computed by solving the dual problem (12), accordingly, the ω and f(x) can be described by:

$$\begin{aligned} \omega &= \sum_{k=1}^{m} [\alpha_k - \alpha_k^* - \sum_{i=1}^{N} \gamma_i p_{ik} \phi(x_k)] \\ f(x) &= \sum_{k=1}^{m} [\alpha_k - \alpha_k^* - \sum_{i=1}^{N} \gamma_i p_{ik} \phi(x_k)] K(x_k, x) + b \end{aligned} \tag{13}$$

and the value of b can be also computed according to the Karush-Kuhn-Tucker (KKT) conditions. This is the formulation of prior knowledge-based support vector regression (PKSVR).

Compared with standard SVR algorithm, PKSVR has the similar expression, so it is convergent and has good generalization ability as well, but importantly it allows for the incorporation of the prior knowledge of special problem, which can be a means of compensating for a usual shortage of training data in practical applications.

SIMULATION EXPERIMENT

The focus of this paper is mainly theoretical. However, in order to illustrate the effectiveness of the proposed formulation, we tested our algorithm on a synthetic example with and without prior knowledge.

We consider the *sinc* function that is expressed as:

$$f(x) = \sin c(x) = \frac{\sin \pi x}{\pi x}$$

which has been extensively used for kernel approximation testing [1, 7, 11, 12].

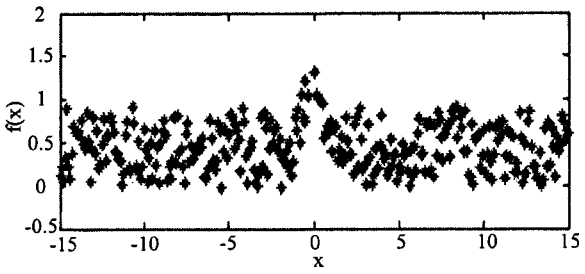

Fig.1: *Sinc* function with white gaussian noise

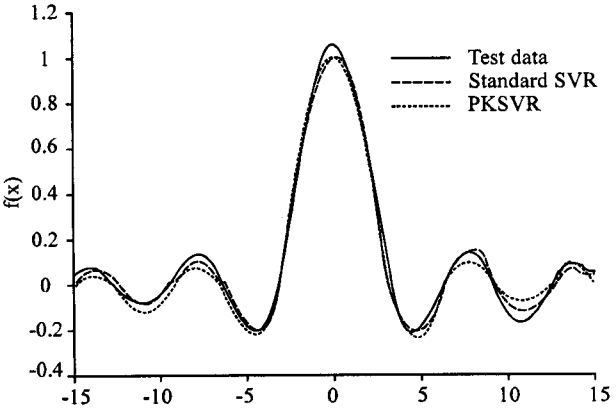

Fig. 2: Comparison of standard SVR and PKSVR

As shown in Fig.1, training data come from *sinc* function with white Gaussian noise for 301 points on the interval . $-15 \le x \le 15$

Fig. 2 shows the regression result of SVR algorithm with and without prior knowledge, i.e. standard SVR algorithm and PKSVR algorithm. The prior knowledge used to approximate the *sinc* function in figure 2 is the following two facts: i) *sinc* function is symmetric about origin, that is,

$x_1 + x_2 = 0 \Rightarrow y(x_1) - y(x_2) = 0$ i) The value $\frac{\sin(\pi/4)}{\pi/4}$

is the minimum of *sinc* on the interval, $-\frac{1}{4} \le x \le -\frac{1}{4}$

that is, $-\frac{1}{4} \le x \le -\frac{1}{4} \Rightarrow f(x) \ge \frac{\sin(\pi/4)}{\pi/4}$

As can be seen from figure 2 that incorporating of prior knowledge into standard SVR is effective and can improve the performance of the SVR.

CONCLUSIONS

In this paper, we have briefly reviewed the classical formulation of support vector regression, and presented a prior knowledge-based support vector regression, which incorporates prior knowledge in the form of inequalities to be satisfied by the desired regression function. The issues that incorporate prior knowledge into standard SVR algorithm when conventional data may be few or not available constitute an interesting topic for future research. Additional future work includes suitable formulation and refinement of prior knowledge and applications to computer vision, soft-sensor modeling and quality control, all of which have prior knowledge available.

REFERENCES

1. Vapnik, V.N., 2001. The Nature of Statistical Learning Theory, 2nd (Edn.,) Springer, New York.
2. Lin. C.F. and S.D. Wang, 2004. Training algorithms for fuzzy support vector machines with noisy data, Pattern Recognition Letters, 25: 1647-1656.
3. Kivinen, J., A. Smola and C. Robert, 2004. Williamson, Online learning with kernels, IEEE Transactions on Signal Processing, 52 : 2165-2176.
4. Fung, G., O.L. Mangasarian and J. Shavlik, 2001. Knowledge-based support vector machine classifiers, Data Mining Institute Technical Report 01-09.
5. Schölkopf, B., P. Simard, A. Smola and V.N.Vapnik, 1998. Prior Knowledge in Support Vector Kernels. In Schölkopf, B., C. Burges and A.Smola, (Ed,) Advanced in Kernel Methods-support Vector learning, MIT Press,
6. Mangasarian, O.L., J.W. Shavlik and E.W. Wild, 2004. Knowledge-based kernel approximation, J. Machine Learning Res., 5 : 1127-1141.
7. Sun, B.Y., 2004. A study of the support vector machine and its applications, Ph.D Thesis University of Science and Technology of China, Hefei.
8. Smola, A. and B. Schölkopf, 2004. A Tutorial on Support Vector Regression, Statistics and Computing, 14: 199-222.
9. Mangasarian, O.L. and D.R. Musicant, 2002. Large scale kernel regression via linear programming, Machine Learning, 46: 255-269.
10. Farag, A.A. and M. Refaat, Mohamed, 2004. Regression using support vector machines: Basic foundations. Technical Report. University of Louisville.
11. Baudat, G. and F.Anouar, 2001. Kerneal-based methods and function approximation, International Joint Conference on Neural Networks, Washington, DC., pp: 1244-1249.
12. Suykens,J., J. Brabanter and L. Lukas, Weighted least squares support vector machines: Robustness and sparse approximation, Neurocomputing, 48 : 85-105.