Support Vector Regression Hybrid Algorithm Based on Rough Set

Jiuying Deng, Qiang Chen
Department of Computer Science
Guangdong Institute of Education
Guangzhou, China
djy1111@126.com; cq_c@gdei.edu.cn

Zongyuan Mao, Xiangjun Gao
College of Automatic Science & Engineering
South China University of Technology
Guangzhou, China
auzymao@scut.edu.cn; xjgao89@yahoo.com.cn

***Abstract*—Support Vector Machine has good generality. Its development for function regressing is not as same as that with fast speed for sample separated. Sequence Minimum Optimizing (SMO) is effective on large samples, and is used to handle the problems with sparse solutions. Considering the power of Rough Set (RS) for handling imprecise data, the datum boundary sought by RS will substitute original inputs as training subset. As the size of both training set and support vectors gained reduce, learning machine can be promoted and favor high quality solutions. Based on rough set and SMO algorithm of regression, a hybrid algorithm (RS-SMO-RA) is presented for function regressing. Only a simple and short module is need to makeup for differentiating boundary sample, and then algorithm RS-SMO-RA can outperform common regression algorithm of SMO. At last, experimental results are displayed with two approaches. There are evaluations of two algorithms implementing and testing.**

***Keywords*—support vector regression, SMO algorithm, boundary sample, rough set**

I. INTRODUCTION

Support Vector Machine (SVM) has been developed in last years, mostly about support vector classifier. The distinct sequential minimal optimization (SMO) algorithm[1] analytically solves the problem of optimizing two variants each time for improving the complexity of SVM training, and is effectively applied in the situation of objects with large datum. There are also the SMO algorithms applied to the case of regression, which contribute to the expansion of SMO algorithm, and the progresses of reinforcing SMO[2-3]. Simultaneity, these works greatly promote the advance of SVM regression explored and applied[4-5].

Support Vector Regression (SVR) has the main features that characterize the maximal margin algorithm: a non-linear function learned by a linear learning machine in a kernel-induced feature space. SVR can be used in pattern recognition with a few samples on dimension independent quantities, such as the number of support vectors. The training complexity is only related to the size of the sample. As the cost function of the original SVM is $O(l^3)$ about time and $O((l+1)^2)$ about space, the performance of SVM learning may get worse in the situation of a large dataset so that the optimization is difficultly to implement. In practice there are frequently very few support vectors, which distribute in the bound closest to the hyper-plane. If training subset only gathers the points fallen in a narrow scope around the convex hull, the support vector regression would get better on the shortened size of sample[6]. Rough set (RS) is a new tool, which can effectively processes imprecise data. The samples within the boundary are nicely collected on the RS methodology, and are formed to replace original inputs as the training subset of SVM. We merge RS and SVR on SMO algorithm, and introduce a hybrid regression RS-SMO-RA in which SVR is improved for pattern recognizing of large datum objects. It is analogy of the classification case that the regression function with an ε-insensitive band is found on dual close-bags $\{D^+=y+\varepsilon, D^-=y-\varepsilon\}$. The decision attribute is directly related to the original output y, so that the routine for producing the boundary set performs shortly and efficiently. In experiment we select a typical function added the stochastic noise as input examples, and contrast the solutions of RS-SMO-RA to that of SMO-RA without RS. The resultant diagrams validate that the RS-SMO-RA algorithm is superiority and simplification over SMO-RA algorithm.

II. SUPPORT VECTOR REGRESSION

As in the classification case the learning algorithm of regression minimizes a convex function and its solution is sparse. There is a training sample T, as follows.

$$T=\{(x_1, y_1),\ldots,(x_l, y_l)\} \in (X, Y)^l \quad (1)$$

Where $x_i \in X=R^n$ is the input element, $y_i \in Y=R$ is the output. The difference from the classification is that the output values are real rather than two discrete values (states). The problem solved relies on training sample and, after learning, is to induce the output y corresponding to a test-point.

A. Support Vector Machine for Regression and Classification

Considering a linear regressor, the function $f(x)$ sought, is linear.

$$y=f(x)=(w \bullet x)+b \quad (2)$$

This function corresponds to a hyper-plane in space $R^n \times R$, with the most simple n=1. The linear regression is a problem

978-1-4244-1674-5/08 /$25.00 ©2008 IEEE

for generating a line, which minimizes error between outputs and known examples.

Definition 1 Suppose a training set T, and $\varepsilon>0$, $y=(w \bullet x)+b$ is called as a hard ε-band hyper-plane if all points in T are included in. It is that the hyper-plane must meet constraints:

$$-\varepsilon \leqslant y_i-((w \bullet x)+b) \leqslant \varepsilon, \qquad i=1,2,\ldots,l. \tag{3}$$

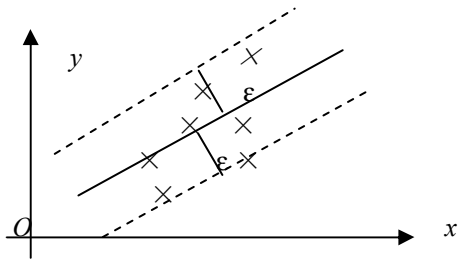


Figure 1. Hard ε-band hyper-plane with n=1

The real line in Fig.1 represents a hyper-plane with n=1, and the scope restricted by two dashed lines is ε-band related to the plane. It is evident that the plane of hard ε-band always presents when the value of ε is enough large and the count of points in training set is finite. The minimum ε_{min} corresponding to being plane of hard ε-band is the solution of optimization in following.

$$\min \quad \varepsilon \tag{4}$$

$$\text{s.t.} \quad -\varepsilon \leqslant y_i-((w \bullet x)+b) \leqslant \varepsilon, \qquad i=1,2,\ldots,l. \tag{5}$$

We form two categories out of T, noted D^+ and D^- respectively as follows;

$$D^+=\{(x_i^T, y_i+\varepsilon)^T, \ i=1, \ ,l\} \tag{6}$$

$$D^-=\{(x_i^T, y_i-\varepsilon)^T, \ i=1, \ ,l\} \tag{7}$$

So, the problem induced is to divide apart two kind of points, both positive and negative, as the classification case.

Theorem 1 Let training set T converted, $(x_{l+i}^T, y_{l+i}-\varepsilon)^T=(x_i^T, y_i-\varepsilon)^T$, i=1,…,$l$, and organize again the sample in following.

$$\{((x_1^T, y_1+\varepsilon)^T;1),\ldots, ((x_l^T, y_l+\varepsilon)^T;1), ((x_{l+1}^T, y_{l+1}-\varepsilon)^T;-1),\ldots, ((x_{2l}^T, y_{2l}-\varepsilon)^T;-1)\} \tag{8}$$

To adapt the same method of a linear separable sample for the maximal margin hyper-plane, the weight vector of which is $(w^T, \eta)^T$, and to solve the optimization problem:

$$\min \ 1/2\|w\|^2+1/2\eta^2 \tag{9}$$

$$\text{s.t.} \ \ z_i((w \bullet x_i)+\eta\,(y_i+z_i\varepsilon)+b) \geqslant 1, \qquad i=1,\ldots,2l \tag{10}$$

$$\text{where, } z_i=\begin{cases} 1, & i=1,\ldots,l; \\ -1, & i=l+1,\ldots,2l; \end{cases} \tag{11}$$

the hyper-plane drawn is $(w^* \bullet x)+\eta^* y+b^*=0$, and after coordinating the linear regress function is derived as: $y=(w \bullet x)+b$, where $w=-w^*/\eta^*$, $b=-w^*/b^*$.

The optimization correspondence of primal problem Eqs.(9)-(10) is in following.

$$\min_\alpha \frac{1}{2}\sum_{i,j=1}^{2l} z_i z_j \alpha_i \alpha_j ((x_i^T, y_i+z_i\varepsilon)^T \bullet (x_j^T, y_j+z_j\varepsilon)^T)-\sum_{j=1}^{2l}\alpha_j \tag{12}$$

$$\text{s.t.} \quad \sum_{i=1}^{2l} z_i\alpha_i = 0 \quad \alpha_i \geqslant 0, \ i=1,\ldots,2l \tag{13}$$

Using the feature space implicitly defined by the kernel $K(x_i, x_j)$, curve instead of line, the hard ε-band method of the maximal margin hyper-plane is popularized to apply in regression as in the classification case. The problem is transferred to the linear regression solved in high-dimension Hilbert space, and suppose the parameter $(\alpha^{(*)} \in R^{2l})$ solve the following optimization problem.

To gain the decision function:

$$f(x)=\sum_i (\alpha_i^*-\alpha_i)K(x_i, x)+b \tag{16}$$

where b is computed for any j with $\alpha_j \in (0, C)$ or k with α_k.

$$b=\begin{cases} y_j-\sum_i(\alpha^*_i-\alpha_i)K(x_i, x_j)+\varepsilon & \alpha_j \in (0, C) \\ y_k-\sum_i(\alpha^*_i-\alpha_i)K(x_i, x_k)-\varepsilon & \alpha_k \in (0, C) \end{cases} \tag{17}$$

B. SMO Algorithm forRegression

The SMO algorithm was initially developed in classification case, undergoes improving and altering, and later is suit to resolve regression either simply or effectively [2]. SMO for regression is summarized as follows.

To alter the appearance of Eq.(14)-(16), let $\lambda_i=\alpha^*_i-\alpha_i$ and $|\lambda_i|=\alpha^*_i+\alpha_i$, where $-C \leqslant \lambda_i \leqslant C$; Assume $k_{ij}=K(x_i, x_j)$ for short, and $k_{ij}=k_{ji}$; The decision function and optimization problem are transferred to Eq.(18)-(19).

$$f(x, \lambda, b)=\sum_i \lambda_i K(x_i, x)+b \tag{18}$$

$$W(\lambda)=\varepsilon\sum_i |\lambda_i|-\sum_i y_i\lambda_i+\frac{1}{2}\sum_{i,j}\lambda_i\lambda_j k_{ij} \tag{19}$$

Which subject to $\sum_i\lambda_i=0$. Suppose two optimized parameters λ_u and λ_v, $s^*=\lambda_u+\lambda_v=\lambda^*_u+\lambda^*_v$, where the symbol "*" indicates values before optimizing; The loop looks for two new KKT violations and optimizes the points until terminating without KKT violations. For the regression the KKT conditions are

$$|y_i-f_i| \begin{cases} <\varepsilon, & \lambda_i=0 \\ =\varepsilon, & -C<\lambda_i \neq 0<C \\ >\varepsilon, & |\lambda_i|=C \end{cases} \tag{20}$$

The popularized SMO of regression is summarized.

Step1: $s^*=\lambda^*_u+\lambda^*_v$, $\eta=k_{uu}+k_{vv}-2k_{vu}$, $\Delta=2\varepsilon/\eta$;
Step2: $\lambda_v=\lambda^*_v+1/\eta\ (y_v-y_u+f^*_u-f^*_v)$;
Step3: $\lambda_u=s^*-\lambda^*_v$;
Step4: if($\lambda_u \bullet \lambda_v<0$)
{ if($|\lambda_v|\geqslant\Delta\wedge|\lambda_u|\geqslant\Delta$)
$\lambda_v=\lambda_v-\mathrm{sgn}(\lambda_v)\bullet\Delta$;
else
$\lambda_v=\mathrm{step}(|\lambda_v|-|\lambda_u|)s^*$; }
Step5: $L=\max(s^*-C,-C)$, $H=\min(C, s^*+C,)$;
Step6: $\lambda_v=\min(\max(L,\lambda_v),H)$;
Step7: $\lambda_u=s^*-\lambda_v$.

b is updated based on any each of both parameters complying to the condition, else average of both , or does nothing.

$$b_u=y_u-f^*_u+(\lambda^*_u-\lambda_u)k_{uu}+(\lambda^*_v-\lambda_v)k_{uv}+b^* \quad (21)$$

$$b_v=y_v-f^*_v+(\lambda^*_u-\lambda_u)k_{uv}+((\lambda^*_v-\lambda_v)k_{vv}+b^* \quad (22)$$

III. Support Vector Machine Regression Based on RS

The support vectors, which determine the decision function, usually distribute over the band-scope around convex hull. Rough set is used to collect points in the band-scope from input sample, and to constitute the sample as training subset of regression learning instead of original input. The performance of machine learning can work better on smaller training set.

A. Knowledge of Rough Set

Fundamental concepts in the theory of knowledge are classifications and categories. The knowledge representation system $S=(U, A)$ may be viewed as a description of a knowledge base[8]. The boundary region brings non-preciseness of sets.

The information system is represented as a decision table: (U, A), $A=C\cup D$, $C\cap D=\varnothing$; Where, C is the set of condition attributes, D is the set of decision attributes. Let $W\subseteq U$ be subsets of U and equivalence relation sets U/a of $a\in A$; Low approximate set of W is defined as $W^{(U/a)^-}=\bigcup_{V\in U/a, V\subseteq W} V$, and also called $S_a(W)$ (supporting subset); Upper approximate set of W is defined as $W^{(U/a)+}=\bigcup_{V\in U/a, V\cap W\neq\phi} V$.

In another words, position region of W is $\mathrm{pos}(W)=W^{(U/a)^-}$, negative region $\mathrm{neg}(W)=U-W^{(U/a)+}$, and boundary region $\mathrm{bn}(W)=W^{(U/a)+}-W^{(U/a)^-}$. U is classified upon the attributes, as $U/a_1, U/a_2, \ldots, U/a_{|A|}$; $\mathrm{pos}(W)$ express a set in which equivalence classes surely belong to sets of W composed of elements in U; $\mathrm{neg}(W)$ is a set in which equivalence classes surely do not belong to sets of W composed of elements in U; and $\mathrm{bn}(W)$ is the set composed of elements that belong to neither $\mathrm{pos}(W)$ nor $\mathrm{neg}(W)$. So the boundary sets of attributes are sets of intercrossing every attribute boundary set, as follows:

$$\mathrm{bn}_A(W)=\mathrm{bn}_{a1}(W)\cap\ldots\cap\mathrm{bn}_{an}(W)\ , \qquad n=|A| \quad (23)$$

For the regression the problem becomes very simple. From Eq.(11) we know that the decision attribute Z relates to single variant y; other elements effect y in Eq.(18) and do not effect z directly. So the boundary set in Eq.(23) maintains an attribute item. When y is continuous y must be discretized. Since the discrete values or partition values may change the size and precision of boundary set, the partition is reasonable to set $\delta\in[\varepsilon, 3\varepsilon]$ based on ε effect.

B. Hybrid Regression Algorithm of SMO and RS

The role of rough set is to generate the boundary set of input sample, which replaces the sample as training subset of regression learning. Subsequently the regression SMO algorithm is used to optimize two parameters. So this is a hybrid regression algorithm of SMO and RS, called RS-SMO-RA, which block-figure is illustrated in Flg.1.

There are two stopping criterions: at first the maximum generation limits the maximum circles, and secondly the number of SV is same in successive two runs, either of them means the circulation terminated.

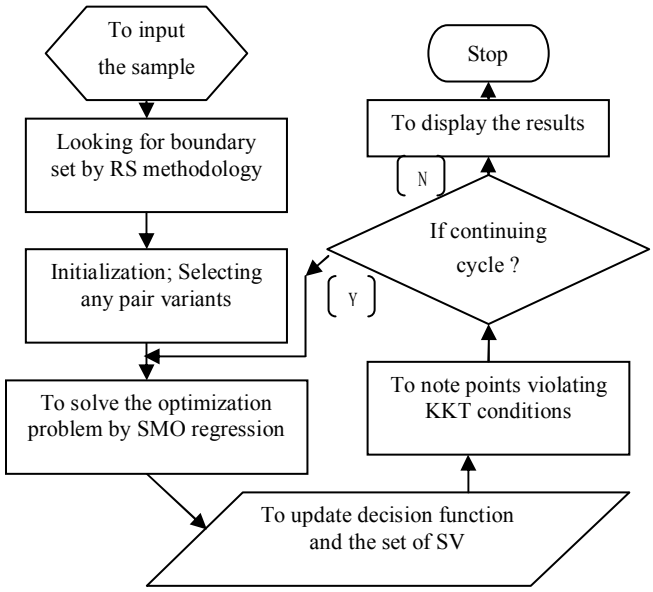


Figure 2 Flowchart of algorithm RS-SMO-RA

IV. The Simulation Results

The RS-SMO-RA algorithm is used to simulate, and the experiment results compare with that of SMO-RA algorithm removing the module for finding boundary set in Fig.1. The original object is the function $\sin(x)/x$, $x\in[-13, 13]$. The points are equably sampled from the function with step $\delta=0.05$, and add stochastic noises $((x_t-0.5)/2)$ to gain the input sample which length is $l_1=521$. Assume $x_t^{(0)}=0.1$; The units of boundary set generated by RS method count $l_2=469$.

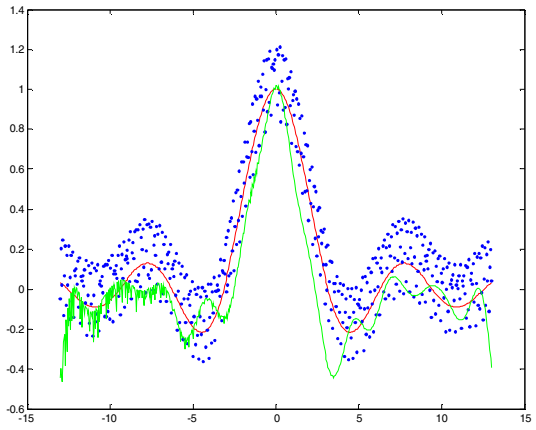
$$x_t^{(k)}=3.78\times x_t^{(k-1)}\times(1-x_t^{(k-1)}) \quad (24)$$

The kernel adopts the form of RBF, $K(x_i, x_j)=\exp(-\|x_i-x_j\|^2)$. To select the parameters ε and C based on the simulation object. In general, ε is so large that there is no reasonable result,

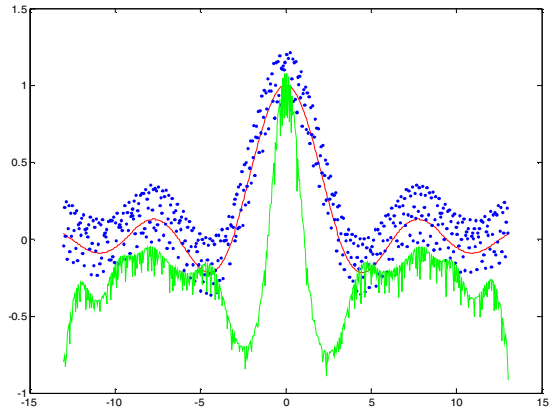
and too small ε would result in the performance degradation of learning machine; the value of *C* is related with the smoothness of received decision function. Two group of regressing results corresponding to different values of ε and *C* are listed in figure 3 and 4 respectively, where, the red curve is the original object graph, the blue scattering dots are the input points with noise, and the green curve is the gained graph of machine learning by training algorithms. The test results in table 1 include main index evaluations: count of SV, actual times of iteration, and time cost.

TABLE I. TESTING RECORDS OF TWO ALGORITHMS

Methods	**RS-SMO-RA**	**SMO-RA**
ε: *C*	.015: .28; .03: .25	.015: .28; .03: .25
Count Of SV	210 187	233 201
Iteration (Gen)	237 237	263 263
Time(s)	14 11	20 16

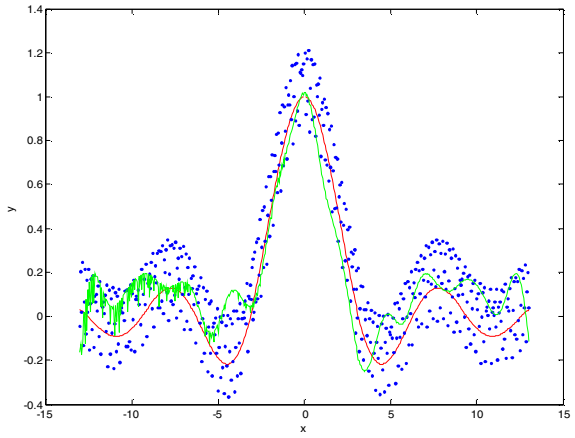


(a) Decision function of machine learning using RS-SMO-RA algorithm

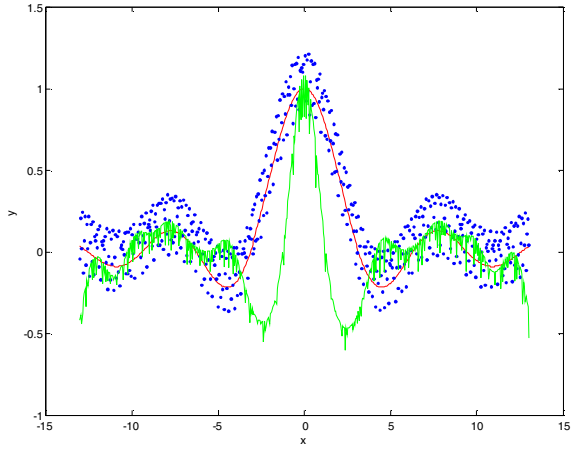


(b) Decision function of machine learning using SMO-RA algorithm

Figure 3. Experiment results of two algorithms with parameters ε=0.015 and *C*=0.28



(a) Decision function of machine learning using RS-SMO-RA algorithm



(b) Decision function of machine learning using SMO-RA algorithm

Figure 4. Experiment results of two algorithms with parameters ε=0.03 and *C*=0.25

Assume the standard deviation of RBF σ=1 in experimenting. As the standard deviation σ changes *C* and ε adjust accordingly. From experiment results it is known that comparing with RS-SMO-RA, SMO-RA implementing collects more support vectors, and draws the sharper and coarser curve. The records of SMO-RA are no good as that of RS-SMO-RA.

V. CONCLUSION

In system identification of large dataset the RS-SMO-RA algorithm has distinct superiority only in cost of short and simple program. Algorithm implementing is independent of the dimension of input sample, and induces the decision function on less SVs for minimum structure risk of SVM regression. The system identifying using this method makes mush progress over using SMO-RA. As the step δ decreases, the ascensive effect enlarges, and the result of machine learning is ever better.

However the result of training algorithm is sensitive to the variety of ε and *C*, which are carefully choose based on actual

object. The choice of ε and *C* merge the flat factor σ of kernel to solve the problem of connected-parameter optimization; otherwise how to select two optimizing points is also important to determine the decision function of system identification. These need to explore the related literatures, and research further.

ACKNOWLEDGEMENTS

This research was supported in part by Guangdong Scientific Technology Planned Sustention Project (No. 2007B010200038) from P.R. China.

REFERENCES

[1] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," Tech. Rep. MSR-TR 98-14, Microsoft Res., Redmond, WA, Apr. 1998.

[2] Gary William Flake, and Steve Lawrence, "Efficient SVM regression training with SMO," Machine Learning. 2002.46. pp.271-290.

[3] YANG Jie, YE Chen-yong, QUAN Yong, and CHEN Nian-yi, "Simplified SMO algorithm for support vector regression," Infrared and Laser Engineering. 2004.33(10). pp.371-375.

[4] S.K.Shevade, S.S.Keerthi, C.Bhattacharyya, and K.R.K.Murthy, "Improvments to the SMO algorithm for SVM regression," IEEE Transactions on neural networks. 2000.9(11). pp.1188-1193.

[5] D. Mattera, F. Palmieri, and S. Haykin, "An explicit algorithm for training support vector machines," Signal Processing Letters, IEEE. 1999.9(6). pp:243 – 245.

[6] ZENG Shao-hua, WEI Yan, DUAN Ting-cai, and CAO Chang-xiu, "Research on algorithm of support vector stepwise regression," Computer Engineering and Applications. 2007.43(8). pp.78-81.

[7] Naiyang Deng, and Yingjie Tian, "New Methodology of Data Mining---Support Vector Machine," Beijing: Science Publisher, 2006.12. pp.77-95.

[8] Wenxiu Zhang, Weizhi Wu, Jiye Liang. Rough Sets Theory and Methodology.[M] Beijing: Science Publisher, 2001.7:24-34.