ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN ITERATIVE LEARNING ALGORITHM BASED ON LEAST SQUARES SUPPORT VECTOR REGRESSION MACHINES

YUPING YUAN, ZENGLONG AN

College of Sciences Heilongjiang Bayi Agricultural University 163319, Heilongjiang, China

ABSTRACT

Aiming at the problem of the large training data set leading to amounts of calculation a sparse approximation algorithm of least squares support vector machine are proposed. Firstly using the thought of matrix block, convert the optimization problem of a Least Squares support vector machine into low - order symmetric positive definite linear systems. Furthermore, use conjugate gradient algorithm which has the number of smaller conditions ,reduce the number of iterations about least squares support vector machine learning process At the same time, we also proved theoretically that the new algorithm is of faster convergence rate. Dramatically improve the speed of the algorithm for learning and prediction. Finally, experimental results show that the algorithm has a very good performance in terms of the accuracy of forecast and speed of training.

Keywords: *Nonlinear Modeling, Least Squares Support Vector Machine, Support Vector Pruning, Conjugate Gradient Method*

1 INTRODUCTION

Support vector machine (SVM), motivated by statistical learning theory in Ref. [1], is a useful tool for classification and regression. It is better to solve the practical problems in the past learning methods. For example, small samples, nonlinear, higher dimensions, over-learning, local minimum point etc. It has many advantages of simple structure, the global optimum and good generalization capability. Its excellent ability to learn makes its wide application in just a dozen years, achieving a series of high-profile research results. It has become a standard tool in the field of machine learning and data mining from the initial linear classification to various pattern classification, regression analysis and time series prediction.(see in ref.[2-3]).But as a new intelligent information processing method researched in recent years, there are still many deficiencies.

Suykens and his associates established Least squares Support Vector Machine in 1999.(see in ref.[4]), which is an improved support vector machine proposed by Suykens, adhering to the basic idea of support vector machines, simplifying its computational complexity. It uses a square of training error instead of relaxation variable, turning the previous optimization of functions constrained by inequality into constrained by the equation. By solving a set of equations, analytical solution of parameter was obtained, avoiding the solution of quadratic programming problem in the dual space of the SVM. It has many advantages such as lowering computational complexity, enhancing solving speed, improving generalization ability. A series of theories and applications of research results were produced based on these advantages of LS-SVM.(see in ref.[5-6]) .

But the disadvantage of LS-SVM is losing the sparse nature of the SVM. As Lagrange multipliers corresponding with a lot of data in the SVM are zero Support Vector is only part of the training set, and just part training data are needed to express SVM model, however, Lagrange multiplier is not zero in LS-SVM No matter what the location of the training point, whether there is noise pollution or not, the role all the training data played in training is the same. So all the training data is a support vector, use all of the data to establish LS-SVM model. Therefore, it is necessary to establish efficient learning algorithm.

Firstly using the thought of matrix block, convert the optimization problem of a Least Squares support vector machine into low - order symmetric positive definite linear systems. Furthermore, use conjugate gradient algorithm which has the number of smaller conditions ,reduce the number of iterations about least squares support vector machine learning process At the same time, we also proved theoretically that the new algorithm is of faster convergence rate. Dramatically improve the speed of the algorithm for learning and prediction. Finally, experimental results show that the algorithm has a very good performance in terms of the accuracy of forecast and speed of training.

2 REGRESSION ALGORITHM OF THE LEAST SQUARE SUPPORT VECTOR MACHINE

Given a training set

$$S=\{(x_i,y_i),x_i\in R^n,y_i\in R,i=1,2,\cdots,l\}$$

x_i which is called an input vector, y_i is known as value of the target corresponding to x_i, l denotes the sample number. The goal of regression problem is to determine the optimal function $f(x)$. Making $f(x)$ forecast unknown input vector correctly with high probability as possible. Regression function has the following form:

$$f(x)=w^T\varphi(x)+b$$

(1)

In the least square support vector machine, the optimization problem corresponding to return problem can be described as

$$\min_{w,b,e} Q(w,b,e)=\frac{1}{2}\|w\|^2+\frac{\gamma}{2}\sum_{i=1}^{l}e_i^2$$

$$s.t\quad y_i=w^T\phi(x_i)+b+e_i(i=1,2,\cdots,l)$$

(2)

Lagrange function is

$$L(w,b,e,\alpha)=\frac{1}{2}\|w\|^2+\frac{\gamma}{2}\sum_{i=1}^{l}e_i^2-\sum_{i=1}^{l}\alpha_i\left[w^T\phi(x_i)+b+e_i-y_i\right]$$

(3)

The best conditions is

$$\begin{cases}\dfrac{\partial L}{\partial w}=0\Rightarrow w-\sum_{i=1}^{l}\alpha_i\phi(x_i)\\ \dfrac{\partial L}{\partial b}=0\Rightarrow\sum_{i=1}^{l}\alpha_i=0\\ \dfrac{\partial L}{\partial e_i}=0\Rightarrow Ce_i-\alpha_i=0\\ \dfrac{\partial L}{\partial \alpha_i}=0\Rightarrow w^T\phi(x_i)+b+e_i-y_i=0\end{cases}\qquad(4)$$

type (4) can be written as the form of a matrix :

$$\begin{bmatrix}I&0&0&-Z^T\\0&0&0&-I\\0&0&CI&-I\\Z&I&I&0\end{bmatrix}\begin{pmatrix}w\\b\\e\\\alpha\end{pmatrix}=\begin{pmatrix}0\\0\\0\\y\end{pmatrix}$$

(5)

where

$$Z=\left[\phi(x_1),\phi(x_2),\cdots,\phi(x_l)\right]^T$$

$$y=[y_1,y_2,\cdots,y_l]^T$$

$$e=[e_1,e_2,\cdots,e_l]^T,\quad\alpha=[\alpha_1,\alpha_2,\cdots,\alpha_l]^T$$

Eliminate w and e ,then get the following equation

$$\begin{bmatrix}0&\vec{1}^T\\\vec{1}&\Omega+C^{-1}I\end{bmatrix}\begin{bmatrix}b\\\alpha\end{bmatrix}=\begin{bmatrix}0\\y\end{bmatrix}$$

（6）

where， $\Omega_{ij}=K(x_i,x_j)$， $B=\Omega+C^{-1}I$,

order $\Phi=\begin{bmatrix}0&\vec{1}^T\\\vec{1}&B\end{bmatrix}$， if matrix Φ can reverse, then

$$\begin{bmatrix}b\\\alpha\end{bmatrix}=\Phi^{-1}\begin{bmatrix}0\\y\end{bmatrix}$$

Solving equations (6) can get regression function

$$f(x)=\omega\cdot\varphi(x)+b=\sum_{i=1}^{l}\alpha_iK(x,x_i)+b$$

(7)

where

$$b=\frac{\overline{1}^TB^{-1}y}{\overline{1}^TB^{-1}\overline{1}}$$

$$\alpha=B^{-1}\left(y-b\overline{1}\right)$$

（8）

define the parameters a,b as regression parameters. type (7) indicates that the key to determining return parameters lie in the calculation of matrix B^{-1}.

3 LS-SVM CONJUGATE GRADIENT ALGORITHM

The sparse nature of the SVM has a strong demonstrating advantage. As for new input, adopt regression function made up of a small amount of support vector, can greatly reduce the amount of calculation. However the biggest drawback of LS-SVM is the loss of the sparse nature .There are two main reasons in Ref. [7]: one is using the quadratic loss function $\sum_{k=1}^{N}e_k^2$ in the objective function of the LS-SVM algorithm ,the other is that the size of the supported value is proportional to error of training point $\alpha_k=\gamma e_k$,

For a new input, All the training samples need to be involved in the operation, which definitely increase the amount of calculation, so in the practical application of the LS-SVM algorithm, it is necessary to carry out sparseness of the solutions of vectors.

3.1 Pruning Algorithm

Because the size of $|\alpha_k|$ reflect the relative importance of each training samples in solution of vector structure. supported value mapping can be got after $|\alpha_k|$ is ranked by descending order, remove training samples with the smaller value of $|\alpha_k|$.Deleting a training sample is equal to make the corresponding support value zero, that is $\alpha_k = 0$.

Thus by constantly removing the training sample which has little impact on decisions or regression function in the training set, then train again new gotten training set, the better sparseness can be got.(see in ref.[8])

Specific steps as the following:

Step 1: Given N number of training samples, use LS-SVM algorithm for training.

Step 2: After descending $|\alpha_k|$, the supported value mapping can be got.

Step 3:Delete training samples with small values in a training set (the number of the deleted samples normally account for 5% in the training set).

Step 4:use LS-SVM algorithm to train newly-gotten training set.

Step 5:Repeat step 2 until the features user defined sharply declined, then algorithm stops.

The pruning algorithm is similar to algorithm of removing hidden nodes used in the neural network, But there's no need for Solving the Inverse Matrix of the Hessian matrix, Only need to make support value with less absolute value in solutions for vector zero, Therefore, the algorithm can be operated simply.

3.2 Conjugate Gradient Algorithm Of Iterative Ls-Svm

Conjugate-Gradient method was first proposed by Hestenes and Stiefle in 1952, it was used for relief of linear equations of Positive Definite Matrix .Based on this, Fletcher and reeves put forward the conjugate gradient method which solved nonlinear optimization problems in 1964. see in ref.[10]. As the conjugate gradient method does not require storage matrix and has advantages of faster convergence rate and secondary termination, conjugate gradient method is now widely used in practical problems. Conjugate gradient method is one of the most effective solution to large linear and nonlinear equations.

Theorem3.1

Set function

$\phi(\cdot): R^n \to R^h, K(x_i, x_j) = \phi(x_i)^T \phi(x_j), y_k \in R, i, j, k = 1 \cdots N$

$C > 0, A \in R^{n \times n}, A_{ij} = K(x_i, x_j) + C^{-1} I$ and

$A \in R^{N \times N}, A_{ij} = y_i y_j K(x_i, x_j) + C^{-1} I$ is a symmetric Positive Definite Matrix

Lemma 3.2

Set λ_{ii} as the eigenvalues of the matrix $K(x_i, x_j)$, if the regular factor meet $0 < C < (\max(|\lambda_{ii}|))^{-1}$, then Matrix $K(x_i, x_j) + C^{-1} I$ is a symmetric Positive Definite Matrix.

Lemma 3.2 gives the range of regular factors C, by the above theoretical analysis, we can conclude : by adjusting the size of the regular factor C, which always cause coefficient matrices corresponded to learning problem as symmetric Positive Definite Matrix. On the other side the conclusion explained the causes why core function non - Positive Definite Matrix in LS-SVM can achieve better learning outcomes.

In (6)，Let

$$A = \begin{bmatrix} 0 & \vec{1}^T \\ \vec{1} & \Omega + C^{-1} I \end{bmatrix}, x = \begin{bmatrix} b \\ \alpha \end{bmatrix}, c = \begin{bmatrix} 0 \\ y \end{bmatrix},$$

Get linear equations of form $Ax = c$

Algorithm 3.3(Conjugate gradient algorithm for program (7))

Step 1:Give an arbitrary starting point $x_0 \in R^n$, calculate $r_0 = c - Ax_0$,let $p_0 = r_0$, ε=0.001, $k = 0$.

Step 2: If $\|r_k\| < \varepsilon$,then stop, Output $x_k \approx x_*$, otherwise go to the next step.

Step 3: Compute

$$\alpha_k = \frac{r_k^T p_k}{p_k^T A p_k}, x_{k+1} = x_k + \alpha_k p_k, r_{k+1} = c - Ax_{k+1},$$

$$\beta_k = -\frac{r_{k+1}^T A p_k}{p_k^T A p_k}, p_{k+1} = r_{k+1} + \beta_k p_k \ .$$

Step 4: Set $k = k + 1$ and go to Step 1.

4 EXPERIMENTAL DATA

4.1 Artificial Data Sets

Realize the above algorithm in Matlab language, and proceed test algorithm 3.3 by a function of one variable in reference [11].We implemented our numerical experiments using Matlab v7.0 on Intel Pentium IV 3.00GHz PC with 512MB of RAM. In the implementation, a function expressed as follows:

$$y = x\sin(4\pi x)e^{1-x^2} + 2x^2\tan(10x)\cos(2\pi x)$$
(9)

Using radial basis functions:

$$K(x, x_i) = \exp(-\|x - x_i\|_2^2 / \sigma^2)$$

Regression modeling accuracy is measured by the following defined approximation error, approximation error function is

$$E = \sqrt{\sum_{i=1}^{N}(y_i - f_i)^2 / \sum_{i=1}^{N}(y_i - \bar{y}_i)^2}, \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

Where, y_i is predicted value, f_i is actual value.

Table 1: The percentage of error by ref.[11] and Algorithm 3.3

Algorithm approximate	σ	C	ε	The number of iterations	error
ref.[11]	2	/	/	500	0.0637
alg.3.3	1	100	0.24	500	0.0603

Table 2:Training time and approximate error on type (9)

sample number	Training time (s)	approximate error (E)
200	2.85	0.0662
400	10.57	0.0663
800	36.57	0.0673

4.2 Uci Dataset

UCI machine learning data set is a standard data set for detecting learning effects, this article uses the following three UCI dataset: machine CPU dataset, Auto-MPG dataset and Boston housing dataset. (see in ref.[12])。

Compare the results of the proposed algorithm in this article with the Results of reference [13]. see Table 4.3.

Use classification algorithm based on function return to classify iris, wine, glass, the above standard test data can be downloaded freely from the machine learning database. Characteristics of Various types of test sample is shown in table 4.4. In the experiment, try to use majority of samples for testing, a small number samples for training, which can increase the speed of training, for three groups of data iris, wine, glass, training samples respectively account for 45.1%,42.7%,44.2% of the total number of samples .the experiment showed that smaller training sample can achieve better classification accuracy. Because there are no structured glass test data and more serious samples imbalances, if use algorithms 3.3, only get 65% correct rate of classification.

Table 3:Training time by Algorithm of " Suykens ", " Kruif " , " Zeng ", " Jiao "and Algorithm 3.3

Data Set	machineCPU	Auto-MPG	Boston housing
#pts	209	392	506
#atr	6	7	13
Suykens	23.03/s	103.42/s	143.9/s
Kruif	36.17/s	402.83/s	625.6/s
Zeng	1734.50/s	969.11/s	3253.1/s
Jiao	4.11/s	27.00/s	39.7/s
Algorithm 3.3	4.03/s	28.53/s	39.62/s

Table 4: Characteristics of test data and results

Data Set	Iris	Wine	Glass
#pts	150	178	214
#atr	4	13	9
#class	3	3	7
training samples	67	76	95
σ	1000	500	200
support vectors	21	32	58
correct rate	0.9900	0.9546	0.6500
accuracy	0.9900	0.9478	0.6491

5 CONCLUSION

As shown in table 4.2, this algorithm 3.3 has makd approximation error smaller ,efficient and short training time , this advantage is particularly evident in the larger problem, training speed increase up to nearly 3 times as fast as before. Improved algorithm requires less support vector ,which lead to a shorter training time, because in calculation we have fully taken into account that constraints produced by non - support vector have effects on target function, enhanced generalization ability of the least squares support vector machines. By theoretical analysis, we have given the set scope of regular factors, in theory, we have given the explanation that LS-SVM model can also achieve better learning outcomes in the case of non - Positive Definite .On this basis, using the conjugate gradient algorithm to build a fast and efficient learning algorithms, theory and numerical Experiments have shown that the iterative LS-SVM conjugate gradient algorithm model is reasonable, simple, faster and easy to implement, so the operating speed of the algorithm can be improved effectively.

ACKNOWLEDGEMENTS

This work was supported by a doctoral subject from the ministry of education (20112305110002) .

This work was supported by the project of science and technology in Heilongjiang Reclamation Bureau (HNK11A-14-07).

REFERENCES:

[1] Cristianini N and Shawe Taylor. An introduction to support vector machines, Cambridge University Press, 2000.

[2] Schlkopf B and Smola A. Learning with kernels, Cambridge MA: MIT Press, 2002.

[3] Naiyang Deng and Yingjie Tian. New methods for data mining-support vector machine, Beijing: Science Press, 2004

[4] J.A. K.Suykens, J. Vandewalle. Least squares support vector machine classifiers. Neural Processing Letters, Vol. 2, No. 9, 1999, pp. 293-300.

[5] Bing Liu, Hongye Shu, An algorithm of Predictive Control Based on least squares support vector machine, Control and decision, Vol. 19, no. 12, 2004, pp. 1439-1442.

[6] J.A. Suykens, J. Vandewalle, Optimal control by least squares support vector machines, Neural Networks, Vol. 14, No. 1, 2001, pp. 23-35.

[7] J.A.K, Suykens, Least Squares Support Vector Machines. NATO-ASI Learning Theory and Practice, Leuven, July 2002

[8] J.A. K.Suykens, L.Lukas, J.Vandewalle. Sparse Approximation Support Vector Macine. IEEE International Symposium on Circuits and Systems, Geneva, Switzerland (Page: 757-760, Year of Publication: 2000)

[9] M.R.Hestenes, E.L.Stiefel. Methods of conjugate gradients for solving linear systems. J Res Nat Bur Standards Sect, Vol. 5, No.49, 1952, pp. 409-436.

[10] R.Fletcher, C.Reeves, Function minimization by conjugate gradients, Computer Journal, 1964，

[11] Fangfang Wu, Yingliang Zhao. Least squares littlewood-paley wavelet Support Vector Machine, information and control, Vol. 34, No.51, 2005, pp.604-609

[12] P.M.Murphu, D.W.Aha, UCI repository of machine learning database, http://www. ics. uci. edu/~mlearn/ MLRepository. Html

[13] Yongping Zhao, Jianguo Sun, A fast and sparse for least squares support vector regression machine. control and decision, Vol. 23, No. 12, 2008, pp. 1347-1352.