Time Series Prediction Using Nonlinear Support Vector Regression Based on Classification

MAO XueMin YANG Jie

Abstract— **In this paper, an introduction of traditional time series prediction model using SVM has been first given, and then followed by description of a new network training algorithm and a nonlinear regression algorithm of support vector machine which are based on classification. Compared with traditional SVM regression algorithm, CSVR algorithm is less sensitive and more robust. It is another advantage that the value of the parameters can be set according to individual situation. More importantly, this method can also escape from over-fitting. Finally, an analysis of this new method has been given to demonstrate the validity of this method.**

Keywords—**SVR (support vector regression); time series; regression algorithm; training algorithm; kernel function**

I. INTRODUCTION

TIME Series indicates a series of data, which are collected in sequence. It is useful for making decision to find an effective way or technology to explore the valuable knowledge hidden in the data.

Traditional times series prediction method use statistic and neural network such as ARMA method and association method[1][2]. On one hand, statistical modeling method requires steady, normality, independence in time series; however, it is not suitable for complex time series. On the Other hand, neural network has a good nonlinear approaching ability, but it also has some other problems. There are some disadvantages of that method, such as easy to over training or lack of training, easy to relapse into local minimization, depending on the design skills too much and so on. This problem has not been solved until 1995, when Cortes and Vapnik introduced a new theory which named support vector machine (SVM) [3]. This method is built on Statistical Learning Theory. Compared with traditional Machine Learning Theory, SVM use the principle of Structural Risk Minimization instead of the principle of Empirical Risk Minimization. SVM synthesize Empirical Risk and Confidence Risk, and have wonderful capacity of generalization [6]. At present, there are some researches on time series prediction using SVM abroad [4]. Unfortunately, these time series prediction using SVM are almost all based on special data; give Artificial Chaotic Sequence Data, for example. In this paper, we try to predict time series data using a new SVM regression algorithm based on classification. Because this new SVM regression algorithm, which based on classification, does not need to build an unknown parameters regression model, it can be used to predict many complex time series and has much better validity. At present, there is no such an application of time series prediction using this new method. In this paper, we analyze the training process and the steps of this algorithm. Then, build model, predict time series data and test its validity.

The rest part of this paper is constructed as follows: In section two, we introduce the time series prediction model based on SVM; In section three, we describe the network training algorithm and regression algorithm using SVM based on classification (CSVR), and combine it to traditional prediction model; In section four, we describe a experiment of time series prediction using this new method on personal in come and its disposition of USA: billions of dollars; SAAR(quarterly); Then, in section five, we analyze and conclude the result of this experiment.

II. REGRESSION AND PREDICTION USING SVM

The traditional method of regression and prediction [3][5], which using SVM, is try to map data $\boldsymbol{x}_i$ on a higher dimension feature space F using a nonlinear map ϕ, and regress on this space. We can use this method to transform the nonlinear problem, which is on lower feature space, to the linear regression problem which is on higher feature space. We can obtain the regression function according to Statistical Learning Theory:

$$f(\boldsymbol{x}) = (w, \phi(\boldsymbol{x})) + b \quad (1)$$

where, $\phi : R^n \to F, w \in F$, (,) denotes inner product, ϕ denotes nonlinear map from R^n space to F space, $\boldsymbol{x} \in R^n$ is weight vector, $w \in F$, b is offset.

Traditional method of solving regression problem is to finding a function f, to minimize its Empirical Risk. The method of SVM regression is to minimize the sum of Empirical Risk and Confidence Risk. It make the prediction model has much better ability of approaching and generalization. In equation (1), $\phi(\boldsymbol{x})$ is known. We can require the estimation values of w and b in equation (1), through functional minimization using sample data $(\boldsymbol{x}_i, Y_i)$ as follow:

$$R_{reg}[f] = \mathrm{R}_{emp}[f] + \lambda \|w\|^2 = \sum_{i=1}^{S} C(e_i) + \lambda \|w\|^2 \quad (2)$$

MAO Xuemin, Management School of Hefei University of Technology, Hefei,Anhui, China (e-mail: maoxuemin@yahoo.com.cn).

Yang Jie, Management School of Hefei University of Technology, Hefei,Anhui, China (e-mail: yangjie129@gmail.com).

International Conference on Computational Intelligence for Modelling Control and Automation,and International Conference on Intelligent Agents,Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)
0-7695-2731-0/06 $20.00 © 2006 IEEE

where, $R_{reg}[f]$ denotes Empirical Risk, $\|w\|^2$ denotes Confidence Risk, $C(e_i)$ denotes the Empirical Loss of model, $C(\cdot)$ denotes Loss Function, $e_i = f(\boldsymbol{x}_i) - Y_i = \hat{Y}_i - Y_i$ is the margin between prediction value and reality value, s is sample size. Because ϕ is steady, $\|w\|^2$ indicates the complexity of the model on higher dimension feature space. The little of the value, the little the Confidence Risk is. λ is the regularize parameter of controlling the compromise of sample training loss and model complexity.

For a given Loss Function, Vapnik introduce an insensitive Loss Function ε, ε can be used to control the width of regression approaching loss, control the number of support vectors and ability of generalization. The little of the value, the higher of the precision, and the much of the support vectors, but the ability of generalization reduced. The Empirical Risk of using this Loss Function is $\mathrm{R}^{\varepsilon}{}_{emp}[f] = \frac{1}{s}\sum_{i=1}^{S}|y - f(x)|_{\varepsilon}$.

Solve the equation (2), get:

$$f(x) = \sum_{i=1}^{S}(\alpha_i - \alpha_i^*)(\phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j)) + b \quad .$$

The Inner-Product Computation on high feature space can be defined as the kernel function of support vector machine: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j))$. We can only get its inner product on higher dimension feature space though computing the kernel function of the variable on lower dimension feature space. Solving this Convex Quadratic Programming problem, we can get its nonlinear map:

$$f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)K(\boldsymbol{x}_i, \boldsymbol{x}_j) + b \quad .$$

From Hilbert-Schmidt Theorem, we know any compute in the Mercer Condition can be regarded as the inner product on higher dimension feature space [3]. Several most common used kernel functions are shown as follows:

1) Polynomial Kernel Function:

$$K(\boldsymbol{x}, \boldsymbol{x}_i) = \left[(\boldsymbol{x} \cdot \boldsymbol{x}_i) + 1\right]^q \boldsymbol{x}_i, \quad t > 0, q \in S$$

2) Gauss Kernel Function:

$$K(\boldsymbol{x}, \boldsymbol{x}_i) = \exp\left\{-\frac{\|\boldsymbol{x} - \boldsymbol{x}_i\|}{2\sigma^2}\right\}, \quad \sigma > 0$$

3) Sigmoid Kernel Function:

$K(\boldsymbol{x}, \boldsymbol{x}_i) = \tanh\left(\nu(\boldsymbol{x} \cdot \boldsymbol{x}_i) + c\right)$, v and c are constants.

III. SVM Regression Algorithm based on Classification(CSVR)

A. Summarize

The notions of traditional regression algorithms are basically compute regression function from sample data directly. These methods can only be used in the situation that the regression models are known. In this section, we describe the support vector regression problem in another point of view. In the new regression algorithm, we use support vector classification (SVC) method which named NonLinear Support Vector Regression Based on Classification (CSVR)[7], instead of the support vector regression method. Compared with traditional support vector regression algorithm, the most advantage is that it can be used in the situation which nonlinear model is unknown, but traditional support vector regression algorithm can not do.

B. CSVR Network Training Algorithm

1) PROBLEM

INPUT: Given a list of input sample $[x_{i1}, x_{i2}, \cdots, x_{in}], i = 1, 2, \cdots S$;

OUTPUT: A support vector network

2) STEPS

STEP1: Divide the data into two styles $C_j = \{Z_+, X_j\}(j = 1, 2)$,

Where, $X_j = (\boldsymbol{x}_1, y_1 + (-1)^j \varepsilon, i = 1, \cdots, S, j = 1, 2)$.

STEP2: Generally, C_1 and C_2 are linearly non-separable, so map the two styles of data into a higher dimension feature space though a specifically function ϕ, and then construct two styles of data on the high dimension feature space $C_j = \{Z_{(-1)^{j+1}}, \phi(X_j), j = 1, 2\}$.

STEP3A: Construct the optimal separating hyper plane H on feature space, construct minimize functional

$$\Psi(w) = \frac{1}{2}|w \cdot w|^2,$$

s.t. $z_i\left[(\phi(\boldsymbol{x}_i) \cdot w) + b\right] \geq 1, i = 1, 2, \cdots, S$, where w is the weight of the optimal separating hyper plane.

STEP3B: Considering the effect of noise, there may be some undivided linear problem in feature space. we can construct a Soft Margin separating hyper plane, ξ_i is slack variable, the positive constant C controls the degree of punishment of the error separation sample.

$$\Psi(w, \xi) = \tfrac{1}{2}|w \cdot w|^2 + C(\sum_{i=1}^{1}\xi_i), \quad \text{s.t.:}$$

$$z_j[(\phi(\boldsymbol{x}_i) \cdot w) + b] \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, 2, \cdots, S.$$

STEP4: The solution of optimization problem can be solved by Functional of Lagrange Saddle Point. At last we get the dual problem of the primal problem:

International Conference on Computational Intelligence for Modelling Control and Automation,and International Conference on Intelligent Agents,Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)
0-7695-2731-0/06 $20.00 © 2006 IEEE

$W(\alpha)=\sum_{i=1}^{S}\alpha_i-\frac{1}{2}\sum_{i=1}^{S}\sum_{j=1}^{S}\alpha_i\alpha_j z_i z_j K(\boldsymbol{x}_i,\boldsymbol{x}_j)$, where $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$ is the kernel function under the condition of Mercer.

STEP5: Solve this dual problem, get a series coefficients of Lagrange Multipliers, generally, there are many items which are zeros, the items which are not zeros are support vectors.

Now, we have constructed a CSVR support vector regression network using standard support vector method of bisection.

C. CSVR Regression Algorithm

1) PROBLEM

INPUT: CSVR support vector regression network and testing vectors $[x_{i1},x_{i2},\cdots,x_{in}],i=1,2,\cdots,\mathrm{N}$;

OUTPUT: $y_i,i=1,2,\cdots,\mathrm{N}$.

2) STEPS

STEP1: Suppose vector $X'=[X,Y]$ is a function of $Y=f(X)$, input this vector into CSVR support vectors networks, then its output would on the optimal separating hyper plane, which is $\sum_{i=1}^{\mathrm{N}}\sum_{j=1}^{\mathrm{S}}\alpha_i z_j K(\boldsymbol{x}_i',\boldsymbol{x}_j)=0$. In this equation, y_i which are contained in $\boldsymbol{x}_i'$ are unknown variables, others are all known. If y_i has been got, the problem could be converted into a regression problem: $y_i=f(x_i),i=1,2,\cdots,\mathrm{N}$, under the condition of $\boldsymbol{x}_i\to y_i$.

STEP2: Because each point, which is the output value of the input variables X', locates on the optimal separating hyper plane, equation (3) can be divided into s equations in one unknown which solve y_i:

$$\sum_{j=1}^{\mathrm{S}}\alpha_j z_j K(\boldsymbol{x}_i',\boldsymbol{x}_j)=0,i=1,2\cdots,N.$$

STEP3: Because of the characteristic of support vector machine, the coefficients of Lagrange Multipliers α_i contain large numbers of zeros. Generally, the number of non-zero $l<<N$, the equation above can be simplified as :

$$\sum_{j=1}^{l}\alpha_j z_j K(\boldsymbol{x}_i',\boldsymbol{x}_j)=0,i=1,2\cdots,N.$$

STEP4: There are many methods can solve y_i, such as the method of Steepest Decent Method. Here we use steepest decent method, let $F(y_i)=\sum_{j=1}^{l}\alpha_i z_i K(\boldsymbol{x}_i',\boldsymbol{x}_i)$ as target function, compute $y_i(j+1)=y_i-\eta\frac{dF(y_i)}{dy_i}$ iteratively to solve y_i, where η denotes Step length.

Now we can build a prediction model in the same way with the traditional SVM prediction. Experiment is shown as follows.

IV. EXPERIMENT

A. Description of Experiment

In this experiment, we use the time series data of personal income in United States (Personal income: Personal Income and Its Disposition: Billions of dollars; SAAR (quarterly)). From 1947 to 2005, each quarter corresponds with a value. Let the data from the first quarter of 1947 to the fourth quarter of 2000 as training sample; let the rest 20 data as testing sample. Original data points are shown in the graphs as red point (we use 0-23.5 to describe each quarter from 1947 to 2005, each point represent a quarter):

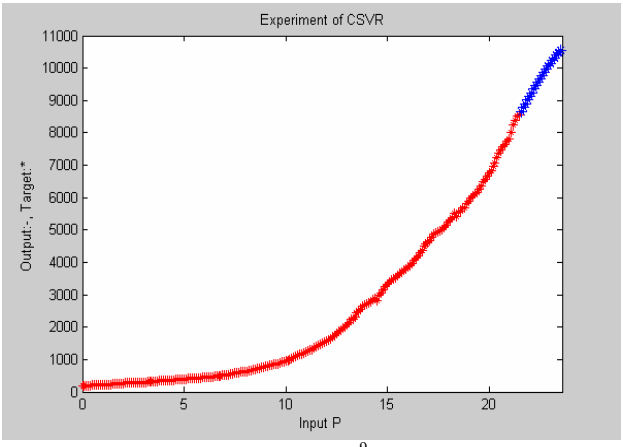

Figure1 σ=6.0*10^9, ε=102.15

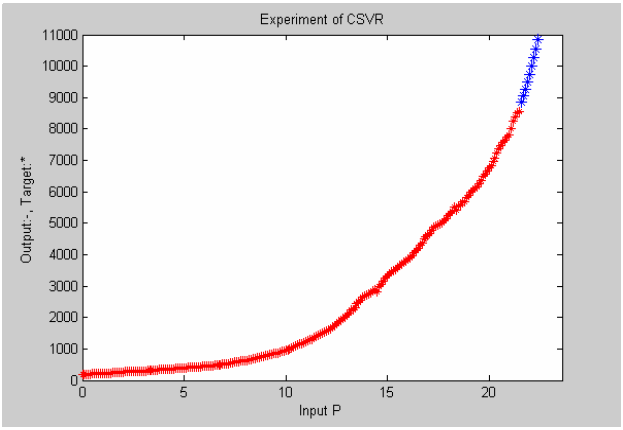

Figure2 σ=6.0*109, ε=100

We use Gauss kernel function in this experiment, choose error separation sample punish parameter C = einsensitive, σ=6.0*10^9, ε=102.15, the results of the experiment is very well (as figure1 show). We can find the prediction results easily, not only its tendency, but also the value of

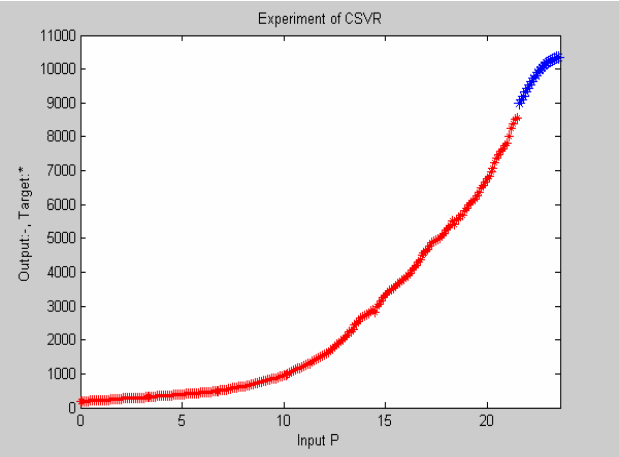

Figure3 σ =6.0*10^9, ε =105

each datum. In the experiment we should adjust the values of δ and ε artificially, for the training sample of support vector machine requests a higher standard of σ and ε, and different δ and ε can get very different result.

B. The Result of Experiment

At the very beginning, we compare the discrepancies between prediction value and reality value of the rest 20 data, and compare the prediction effect of the difference between δ and ε. The results are shown in Table1 and Table2 as follows.

Table1 when σ =6.0*10^9, change the value of ε, the results are as follows (as show in Figure1, Figure2 and Figure3)

Note: P.E. is the abbreviation of Predictive Value.

Quarter		01 / 1	01 / 2	01 / 3	01 / 4	02 / 1	02 / 2	02 / 3	02 / 4	03 / 1	03 / 2
P. E.	ε =105	8983	9099	9212	9322	9429	9532	9631	9725	9814	9897
	ε =102.15	8654	8774	8892	9010	9125	9239	9351	9461	9569	9674
	ε =100	8854	9055	9269	9497	9739	9997	10269	10557	10862	11182
Reality Value		8688.7	8719.9	8733.1	8754.8	8814.7	8892.0	8895.4	8925.5	9013.7	9118.6
Quarter		03 / 3	03 / 4	04 / 1	04 / 2	04 / 3	04 / 4	05 / 1	05 / 2	05 / 3	05 / 4
P. E.	ε =105	9974	10045	10109	10166	10216	10257	10291	10315	10332	10339
	ε =102.15	9776	9875	9971	10065	10154	10240	10323	10402	10476	10547
	ε =100	11518	11870	12237	12619	13015	13424	13845	14277	14719	15169
Reality Value		9215.4	9328.7	9484.8	9614.3	9729.2	10024.8	10073.4	10185.7	10250.4	10483.7

Table2 when choose ε =102.15, change the value of σ, the results are as follows (the figures are omitted here）

Quarter	01 / 1	01 / 2	01 / 3	01 / 4	02 / 1	02 / 2	02 / 3	02 / 4	03 / 1	03 / 2
σ =1.0*10^8	8645.0	8756.9	8865.2	8969.4	9068.9	9163.2	9252.0	9334.6	9410.7	9479.9
σ =1.0*10^{11}	11754	11874	11992	12109	12225	12339	12451	12561	12668	12773
Quarter	03 / 3	03 / 4	04 / 1	04 / 2	04 / 3	04 / 4	05 / 1	05 / 2	05 / 3	05 / 4
σ =1.0*10^8	9541.8	9596.1	9642.3	9680.4	9709.9	9730.7	9742.6	9745.5	9739.3	9724.0
σ =1.0*10^{11}	12876	12975	13071	13164	13254	13340	13423	13501	13576	13646

C. Discussion of the Result

(1) C, which is the Error Dividing Sample Punish Parameter, used to split the difference between the proportion of error dividing sample and algorithm complex degree. C should be chosen by researchers, its value is stochastically. Unfortunately, it is different to decide whether it is good or not. By far, the problem which is the elimination of C's randomicity and employing some method to obtain the optimal value automatically instead, has not been solved theoretically.

(2) When the Insensitive Loss Function ε is introduced, the SVM is extended to regression estimation of a nonlinear system, which demonstrates an excellent learning performance. The regression estimation based on SVM method can approach any nonlinear function with an under-controlling precision, at the same time it has the excellent performance of global optimum and the ability of generalization. The SVM controls the precision of regression estimation though ε, but it is still uncertain that how can get an expected estimation precision though changing ε.

(3) The performance of prediction curve is decided by kernel function. At present, there are several kinds of kernel functions (in section two) and their parameters are all selected by our experience, and it's a limitation. In different problem field, kernel functions would have different format and parameters, so it must introduce specific field knowledge and choose kernel function based on the characteristic of data sequence. But at present there is no good method to solve the problem of choosing kernel function.

(4) When a certain kernel function is chosen, the selection of the value σ plays a vital role for the prediction effect of the curve. Give this experiment, for example, we choose Gauss kernel function, if the kernel width we used is a little big, we may get a smoothness prediction function, so the validity of fitting is a little bad, and the ability of generalization is also very bad. On the contrary, if the kernel width we used is a little small, we may get much sharper Gauss function curve, and each point of the sample is the peak value of the Gauss function, and the ability of generalization is also very

0-7695-2731-0/06 $20.00

bad. So, we should choose the form and parameter according to individual problem.

V. Conclusion and Future Work

As an application method which is developed on Statistical Learning Theory, support vector machine mainly use the principle of Structural Risk Minimization instead of the principle of Empirical Risk Minimization. When applied on regression problem, support vector regression considers curve smooth and error degree synthetically, so it improves the generalization ability. In this paper, we use a new support vector regression algorithm based on classification to predict time series data. Compared with traditional support vector regression method, the support vector regression method based on classification can be used to regress on the condition that the model is unknown forehand, and because it is also a support vector method itself, the efficiency can be guaranteed. Further research include: multi-dimension time series data prediction, and try to solve multi-dimension time series prediction problem using Econometrics model.

References

[1] YiMin Xiong, Di-Yan Yeung, Time series clustering with ARMA mixtures, Pattern Recognition 37(2004) 1675-1689

[2] Ildar Batyrshin, Raul Herrera-Avelar, et al., Association Network in Time Series Data Mining, NAFIPS 2005-2005 Annual Meeting of the North American Fuzzy Information Processing Society.

[3] Vapnik V N. The Nature of Statistical Learning Theory [M], New York: Springer-Verlag, 1995

[4] Sayan Mukherjee, Edgar Osuna, Federico Girosi, Nonlinear Prediction of Chaotic Time Series Using Support Vector Machines, IEEE NNSP'97, Amelia Island, FL, 24-26 Sep., 1997

[5] Vapnik V N, Golowich S E, Smola A. Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing[A], San Mateo, CA: NIPS'8, 1996

[6] Tian Xiang, Deng Feiqi, Application of Accurate Online Support Vector Regression in Stock Market Index Forecasting, Computer Engineering November 2005

[7] Ye Ning, Liang Zuopeng, Dong Yisheng, Wang Huoli, SVM Nonlinear Regression Algorithm, Computer Engineering, October 2005

Author Profiles

MAO Xuemin(1973-), Male, Associate Professor, Management School of Hefei University of Technology, Research interests Decision Support System, Artificial Intelligence.

YANG Jie(1983-),Male, Student of Management School of Hefei University of Technology, Hefei, China, Research interests Information Management and Information System.

International Conference on Computational Intelligence for Modelling Control and Automation,and International Conference on Intelligent Agents,Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)
0-7695-2731-0/06 $20.00 © 2006 IEEE