SUPPORT VECTOR REGRESSION FOR FINANCIAL TIME SERIES FORECASTING

Wei Hao, Songnian Yu
School of Computer Engineering and Science, Shanghai University, Shanghai, China; Email: haowei@graduate.shu.edu.cn

Abstract: Recently, Support Vector Regression (SVR) has been a popular tool in financial time series forecasting. This study deals with the application of Support Vector Regression in stock composite index forecasting. A preprocessing method for accelerating support vector regression training is presented in this paper. Then we propose a method of support vector regression by modifying the regularized risk function. A data set from Shanghai Stock Exchange is used for the experiments to test the validity of our methods.

Key words: Support Vector Regression, Financial Time Series forecasting

1. INTRODUCTION

Stock market prediction is regarded as one of the most challenging applications of modern time series forecasting. Over the past decade, neural networks have been successfully used for modeling financial time series [1,2]. A large number of successful applications have shown that ANN can be a very useful tool for time-series modeling and forecasting [3]. The early days of these studies focused on application of ANNs to stock market prediction [4,5]. Recent research tends to hybridize several artificial intelligence techniques [6,7]. However, ANN has a difficulty in explaining the prediction results due to the lack of explanatory power, and suffers from difficulties with generalization because of overfitting.

Based on Statistical Learning Theory, Support Vector Machine (SVM), was first developed by Vapnik [8,9]. It has become a hot topic of intensive

Please use the following format when citing this chapter:
Hoa, Wei, Yu, Songnian, 2006, in International Federation for Information Processing (IFIP), Volume 207, Knowledge Enterprise: Intelligent Strategies In Product Design, Manufacturing, and Management, eds. K. Wang, Kovacs G., Wozny M., Fang M., (Boston: Springer), pp. 825-830.

study due to its successful application in classification tasks[10,11] and regression tasks[12,13], specially on time series prediction[14] and financial related applications[15]. This paper focuses on the application of Support Vector Machine in regression tasks to make a new attempt to stock composite index forecasting. A novel method for accelerating support vector regression (SVR) training based on a measurement of similarity among samples is presented in this paper. Then, based on the prior knowledge that in the non-stationary financial time series the dependency between input variables and output variable gradually changes over the time, specifically, the recent past data could provide more important information than the distant past data. We modified the regularized risk function in support vector regression, whereby the recent ε -insensitive errors are penalized more heavily than the distant ε -insensitive errors.

2. PREPROCESSING TO REDUCE TRAINING DATA

From the implementation point of view, training SVM is equivalent to solving a linearly constrained quadratic programming (QP) problem with the number of variables equal to the number of training data points. However, with the great increment of training data in size, the memory space for storing the kernel matrix will increase with the level $O(N^2)$, where N is the number of the training data. Hence, designing effective SVM training algorithms for large data sets will be practically significant.

A novel method based on the similarity measurement, for reducing training data to accelerate support vector machines training, is proposed here. The notion of similarity, as a measurement of the approaching degree between two samples, is introduced to select new training data. The similarity function S is defined as:

$$S(x,y) = f\left(\frac{1}{\|x-y\|^2}\right) = f\left(\frac{1}{\sqrt{\sum_{i=1}^{M}(x_i - y_i)^2}}\right) \tag{2.1}$$

those data points, which are close to some referenced data point according to the similarity function, will be discarded because they can be replaced by the special data point without great influence on the final solution[16]. The division of data groups will affect generalization performance and it can be achieved by clustering. In order to reduce possible errors from clustering phase, we designed a novel clustering algorithm based on K-means and t-test,

in which after finding the nearest cluster for a data point, we add a hypothesis testing to verify the statistical significance of correlation between them.

3. THE MODIFIED SUPPORT VECTOR REGRESSION

In the field of financial time series forecasting, numerous studies show that the relationship between input variables and output variable gradually changes over time, and recent data could provide more information than distant data. Therefore, it is advantageous to give more weights on the information provided by the recent data than that of the distant data based on this prior knowledge[17,18]. In the light of this characteristic, we introduced a novel approach whereby more weights are given to the recent ε-insensitive errors than the distant ε-insensitive errors in the regularized risk function. The regularized term in the regularized risk function is retained, regardless of the empirical error.

In Support Vector Regression, the empirical risk function has equal weight C to all the ε-insensitive errors between the predicted and actual values. The regularization constant C determines the trade-off between the empirical risk and the regularized term. Increasing the value of C, the relative importance of the empirical risk with respect to the regularized term grows. For illustration, the empirical risk function is expressed as:

$$E_{svr} = C\sum_{i=1}^{l}(\zeta_i + \zeta_i^*) \tag{3.1}$$

In our approach, instead of a constant value, the regularization constant C adopts a weight function:

$$E_{svr} = \sum_{i=1}^{l} C_i(\zeta_i + \zeta_i^*) \tag{3.2}$$

$$C_i = w(i)C \tag{3.3}$$

Where $w(i)$ is the weight function satisfying $w(i) > w(i-1), i = 2,...,n$. As the weights will incline from the distant training data points to the rent

training data points, C_i will give more weights on the more recent training data points.

4. EXPERIMENTS

Experiments have been conducted to illustrate effectiveness of the presented method. A data set from Shanghai Stock Exchange was used for the experiments. We selected daily closing prices of stock composite index of shanghai stock exchange between Jan.21, 2002 and Sep. 27, 2002. There are totally 140 data points. 100 data points in front of the data series may be used as training data sets and the rest 40 data points as testing data. The experiments will emphasize on verifying the performance of the preprocessing method and evaluating the effectiveness of the modified SVR.

The experiments first emphasize on verifying the effectiveness of preprocessing method for reduction training data. Table 1 summarizes the training and testing results. In practical applications, a suitable similarity threshold may be selected according to specific requirement of speed and accuracy. The prediction performance is evaluated using two standard errors; mean absolute error (MAE) and root mean squared error (RMSE), there are defined as the following, respectively:

$$MAE = \frac{1}{l} * \sum_{i=1}^{l} abs(y_i - d_i), \tag{4.1}$$

$$RMSE = \sqrt{\frac{1}{l}\sum_{i=1}^{l}(y_i - d_i)^2}, \tag{4.2}$$

Where, y_i, d_i denote the predicted result and the measured value, respectively.

Table 1. The training and testing results for different similarity thresholds

S	CPU time (Second)	No. of training vectors	No. of Support vectors	MAE	RMSE
0.2	45.3	83	23	0.8281	0.8791
0.1	18.6	79	20	0.9194	1.0275
0.05	5.4	51	18	1.6259	2.1851
0.03	2.8	37	19	1.5343	1.9146
0.02	1.9	22	16	3.0517	3.6579
0.01	0.5	10	10	32.5634	38.1064

We use general RBF as the kernel function. There are two parameters while using RBF kernels: C and γ. It is not known beforehand which C and γ are the best for one problem; consequently some kind of model selection (parameter search) must be done. The goal is to identify them so that the classifier can accurately predict unknown data (i.e., testing data). Note that it may not be useful to achieve high training accuracy (i.e., classifiers accurately predict training data whose class labels are indeed known). Therefore, a common way is to separate training data to two parts of which one is considered unknown in training the classifier. Then the prediction accuracy on this set can more precisely reflect the performance on classifying unknown data. An improved version of this procedure is cross-validation.

We use a grid-search on C and γ using cross-validation. Basically pairs of (C, γ) are tried and the one with the best cross-validation accuracy is picked. We found that trying exponentially growing sequences of C and γ is a practical method to identify good parameters (for example, $C = 2^{-5},...,2^{15}$, $\gamma = 2^{-15},2^{-13},...,2^{3}$).

Figure 1 illustrates the predicted and actual values. The dotted line is the actual value. The smoothing line is the predicted value of our method. From the figure, it can be observed that the method forecast more closely to the actual values in most of the time period. The experimental results show that this method is more effective and efficient in forecasting stock composite index than the standard SVR and traditional time series forecasting models.

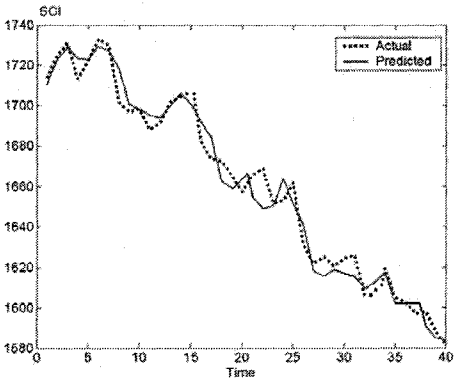

Figure 1. Forecasting result comparison

5. CONCLUSION

In this research, a preprocessing method for accelerating SVR training and a modified SVR are proposed. The objective of this research is also to examine the feasibility of the methods in stock composite index forecasting and to improve the accuracy of SVR in terms of parameters selection. By using our methods, high prediction performance is achieved.

6. REFERENCES

1. W. Cheng, L. Wanger, Forecasting the 30-year US treasury bond with a system of neural networks. *Journal of Computational Intelligence in Finance*, 4 (1996),pp.10 16.
2. R. Sharda, R. Patil. A connectionist approach to time series prediction: an empirical test. *Neural Networks in Finance and Investing*, Chicago: Probus Publishing ,1994, pp.451 464.
3. G. Zhang, B.E. Patuwo, M.Y. Hu, Forecasting with artificial neural networks: the state of the art, *Int.J. Forecasting* 14 (1998) ,pp. 35 62.
4. H. Ahmadi, Testability of the arbitrage pricing theory by neural networks, *Proceedings of the International Conference on Neural Networks*, San Diego, CA, 1990, pp.1385 1393.
5. J.H. Choi, M.K. Lee, M.W. Rhee, Trading S& P 500 stock index futures using a neural network, *Proceedings of the Annual International Conference on Artificial Intelligence Applications on Wall Street*, New York, 1995, pp.63 72.
6. Y. Hiemstra, Modeling structured nonlinear knowledge to predict stock market returns, *Chaos & Nonlinear Dynamics in the Financial Markets: Theory, Evidence and Applications*, Irwin, Chicago, IL, 1995, pp.163 175.
7. I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, *Morgan Kaufmann Publishers*, San Francisco, CA, pp.1999.
8. V.N.Vapnik. The Nature of Statistical Learning Theory. *Springer*, New York, pp.1995.
9. V.N.Vapnik. Statistical Learning Theory. Wiley, New York, 1998.
10. Edgar Osuma and Robert Freund and Federico Girosi. Support Vector Machines: Training and Applications. *AIM-1602*, MIT, 38, 1997.
11. J.C.Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2),pp.121-167, 1998.
12. V.N.Vapink, S.Golowich and A.Smola. Support vector method for function approximation, *regression estimation and signal processing*.
13. A.Smola and B.Scholkopf. A Tutorial on Support Vector Regression. 1998, *Technical Report NeuroCOLT NC-TR-98-030.*
14. K.R.Muller, A.Smola, G.Ratsch, B.Scholkopf, J.Kohlmorgen and V.N.Vapnik. Predicting time series with support vector machines. *ICANN*, 999-1004, 1997.
15. T.B.Trafalis and H.Ince. Support vector machine for regression and applications to financial forecasting. *IJCNN 2000*, 348-353.
16. W.Wang, Z.Xu. A heuristic training for support vector regression. *Neurocomputing* 61 (2004) 259-275.
17. Francis E.H.Tay, L.J.Cao. Modified support vector machines in financial time series forecasting. *Neurocomputing* 48 (2002) 847-861.
18. N.D. Freitas, M. Milo, P. Clarkson, Sequential support vector machines, *Proceedings of the IEEE Signal Processing Society Workshop*, Madison, WI, USA, 1999, 31 40.