Outliers Treatment in Support Vector Regression for Financial Time Series Prediction

Haiqin Yang, Kaizhu Huang, Laiwan Chan, Irwin King, and Michael R. Lyu

Department of Computer Science and Engineering
The Chinese University of Hong Kong
Shatin, N.T. Hong Kong
{hqyang,kzhuang,lwchan,king,lyu}@cse.cuhk.edu.hk

Abstract. Recently, the Support Vector Regression (SVR) has been applied in the financial time series prediction. The financial data are usually highly noisy and contain outliers. Detecting outliers and deflating their influence are important but hard problems. In this paper, we propose a novel "two-phase" SVR training algorithm to detect outliers and reduce their negative impact. Our experimental results on three indices: Hang Seng Index, NASDAQ, and FSTE 100 index show that the proposed "two-phase" algorithm has improvement on the prediction.

1 Introduction

Recently, due to the advantage of the generalization power with a unique and global optimal solution, the Support Vector Machine (SVM) has attracted the interest of researchers and has been applied in many applications, e.g., pattern recognition [1], and function approximation [8]. Its regression model, the Support Vector Regression (SVR), has also been successfully applied in the time series prediction [5], especially in the financial time series forecasting [2]. This model, using the ε-insensitive loss function, can control the sparsity of the solution and reduce the effect of some unimportant data points. Extending this loss function to a general ε-insensitive loss function with adaptive margins has shown to be effective in the prediction of the stock market [9, 10].

In modelling the financial time series, one key problem is its high noise, or the effect of some data points, called outliers, which differ greatly from others. Learning observations with outliers without awareness may lead to fitting those unwanted data and may corrupt the approximation function. This will result in the loss of generalization performance in the test phase. Hence, detecting and removing the outliers are very important. Specific techniques, e.g., a robust SVR network [4] and a weighted Least Squares SVM [6] have been proposed to enhance the robust capability of SVR. These methods would either involve extensive computation or would not guarantee the global optimal solution.

In this paper, we propose an effective "two-phase" SVR training algorithm to detect outliers and reduce their effect for the financial time series prediction. The basic idea is to take advantage of the general ε-insensitive loss function with a non-fixed margin, which can reduce the effect of some data points by enlarging the ε-margin width.

The paper is organized as follows. We introduce the SVR with a general ε-insensitive loss function and state the method of detecting and reducing outliers in Section 2. We report experimental results in Section 3. Lastly, we conclude the paper in Section 4.

2 Outliers Detection and Reduction in Support Vector Regression

In this section, we first introduce the Support Vector Regression (SVR) in the time series prediction. We then propose a general ε-insensitive loss function for applying the adaptive margins. Next, we describe our method to detect the outliers and reduce their influence.

2.1 Support Vector Regression for Time Series Prediction

Time series data can be abstracted as $(\mathcal{X}, \mathcal{Y})$ pairs, where $\mathcal{X} \in \mathbb{R}^d$ denotes the space of input patterns, $\mathcal{Y} \in \mathbb{R}$ corresponds to the target value. Usually, the sample is finite and observed in a successive time interval. An N-instance sample series is described as $(\mathcal{X}, \mathcal{Y}) = \{(\mathbf{x}_t, y_t) |\ \ \mathbf{x}_t \in \mathbb{R}^d,\ \ y_t \in \mathbb{R},\ \ t = 1, \ldots, N\}$. In the financial time series, it may be assumed that all the information can be condensed in the price. Hence, y_t usually represents the price at time t and $\mathbf{x}_t$ represents the p-previous days' prices as $\mathbf{x}_t = (y_{t-p}, \ldots, y_{t-1})$. To analyze this series, one may evaluate a function, f,

$$y_t = f(\mathbf{x}_t) + \sigma_t,$$

from the given N-instance sample series, where σ_t is the noise at time t. The SVR is a currently popular technique to learn the data with good generalization [7, **?**].

Typically, the SVR estimates a linear function

$$f(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b, \tag{1}$$

in a feature space, $\mathbb{R}^f$, by minimizing the following regression risk:

$$R_{reg}(f) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} l(f(\mathbf{x}_i) - y_i), \tag{2}$$

where the superscript T denotes the transpose, ϕ is a mapping function in the feature space, and b is an offset in $\mathbb{R}$. The term $\frac{1}{2}\mathbf{w}^T\mathbf{w}$ is a complexity term determining the flatness of the function in $\mathbb{R}^f$, C is a regularized constant, and l is a cost function.

Generally, the ε-insensitive loss function is used as the cost function [7]. This function does not consider data points in the range of ε-margin, i.e, $\pm\varepsilon$. It can therefore reduces the effect of those data points lying in the ε-margin to the approximation function and controls the sparsity of solution. The ε-insensitive loss function is defined as $l_\varepsilon(f(\mathbf{x}) - y) = \max(|y - f(\mathbf{x})| - \varepsilon, 0)$.

2.2 SVR with a General ε-insensitive Loss Function

In the above, the ε-margin is fixed and symmetrical. This setting may lack the flexibility to efficiently model the volatility of the stock market and can not prefer one-side prediction. In order to overcome these problems, we propose a general ε-insensitive loss function. This function divides the margin into two separate parts, up margin, ε^u, and down margin, ε^d, with each part changing adaptively as formulated below:

$$l_{\varepsilon_2}(f(\mathbf{x}_i) - y_i) = \begin{cases} 0, & \text{if } -\varepsilon_i^d < y_i - f(\mathbf{x}_i) < \varepsilon_i^u \\ y_i - f(\mathbf{x}_i) - \varepsilon_i^u, & \text{if } y_i - f(\mathbf{x}_i) \geq \varepsilon_i^u \\ f(\mathbf{x}_i) - y_i - \varepsilon_i^d, & \text{if } f(\mathbf{x}_i) - y_i \geq \varepsilon_i^d \end{cases} \tag{3}$$

The main contribution of proposing this loss function is that we can adopt adaptive margin with non-fixed and asymmetrical characteristics. This would benefit the stock market prediction, e.g., reflecting the volatility of the stock market or avoiding the down side risk.

Minimizing the regression risk of (2) with the cost function of (3) by the Lagrange method, we obtain the following Quadratic Programming (QP) problem:

$$\min_{\alpha,\alpha^*} \ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(\mathbf{x}_i,\mathbf{x}_j)+\sum_{i=1}^{N}(\varepsilon_i^u-y_i)\alpha_i+\sum_{i=1}^{N}(\varepsilon_i^d+y_i)\alpha_i^*,$$

$$\text{s.t.} \ \sum_{i=1}^{N}(\alpha_i-\alpha_i^*)=0, \quad \alpha_i, \ \alpha_i^* \in [0,C], \quad i=1,\ldots,N,$$

where α_i and α_i^* are the corresponding Lagrange multipliers used to push and pull $f(\mathbf{x}_i)$ towards the outcome of y_i, respectively. $K(\mathbf{x}_i,\mathbf{x}_j)=\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$, the inner product of the mapping function, is the kernel function which satisfies the Mercer's condition.

The above QP problem has a similar form to the original QP problem in the SVR and can be easily implemented or solved by e.g., a commonly used SVM library, LIBSVM [3]. After solving the above QP problem, we obtain the corresponding Lagrange multipliers α_i and α_i^*, and the weight $\mathbf{w}=\sum_{i=1}^{N}(\alpha_i-\alpha_i^*)\phi(\mathbf{x}_i)$; therefore, we get the approximation function as $f(\mathbf{x})=\sum_{i=1}^{N}(\alpha_i-\alpha_i^*)K(\mathbf{x},\mathbf{x}_i)+b$, where the offset b is calculated by exploiting the Karush-Kuhn-Tucker (KKT) conditions (details in [3]).

2.3 Outliers Detection and Reduction

From the KKT conditions, we have

$$\begin{aligned}\alpha_i(\varepsilon_i^u+\xi_i-y_i+f(\mathbf{x}_i))&=0, \quad i=1,\ldots,N,\\ \alpha_i^*(\varepsilon_i^d+\xi_i^*+y_i-f(\mathbf{x}_i))&=0, \quad i=1,\ldots,N,\end{aligned} \tag{4}$$

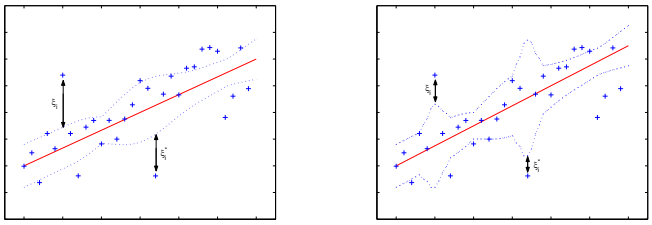
and

$$\begin{aligned}(C-\alpha_i)\xi_i&=0, \quad i=1,\ldots,N,\\ (C-\alpha_i^*)\xi_i^*&=0, \quad i=1,\ldots,N,\end{aligned} \tag{5}$$

where ξ_i and ξ_i^* are slack variables used to measure the error of up side and down side, respectively (see Fig. 1(a)).

The KKT conditions in (5) indicate that if $\alpha_i \in [0,C)$, then $\xi_i=0$; likewise for α_i^* and ξ_i^*. This means that the corresponding data points lie in, or on the ε-margin, i.e., either the ε^u-margin or the ε^d-margin, but not both. Moreover, for $\alpha_i=C$ or $\alpha_i^*=C$, we have

$$\begin{aligned}\xi_i&=y_i-f(\mathbf{x}_i)-\varepsilon_i^u, \ \forall\alpha_i=C, \quad i=1,\ldots,N,\\ \xi_i^*&=f(\mathbf{x}_i)-y_i-\varepsilon_i^d, \ \forall\alpha_i^*=C, \quad i=1,\ldots,N.\end{aligned}$$



(a) (b)

Fig. 1. An illustration of detecting and reducing the effect of outliers in the feature space.

The above formulae show that increasing ε_i^u and ε_i^d will decrease the corresponding ξ_i and ξ_i^* in the same constructed function f. This will therefore reduce the error caused by the corresponding data points. In addition, if we fulfill the objective in (2), a normal point or a non-outlier will not contain a very large error, i.e., ξ_i or ξ_i^*. Based on these observations, we propose the criterion of detecting outliers, i.e., if ξ_i or ξ_i^* is larger than a threshold, the corresponding data point would be an outlier and we could enlarge the corresponding margin width to reduce its effect (see Fig. 1(b)). This motivates us to propose the following "two-phase" procedure:

Phase 1: Train the SVR model with the ε_i^u and ε_i^d margin setting.

Phase 2: Detect and reduce the effect of the outliers. If $\xi_i > \tau\varepsilon_i^u$, $\varepsilon_i^{u'} = \tau\varepsilon_i^u$; similarly we have $\varepsilon_i^{d'} = \tau\varepsilon_i^d$ for $\xi_i^* > \tau\varepsilon_i^d$. Re-train the SVR model by using the updated margin setting, $\varepsilon_i^{u'}$ and $\varepsilon_i^{d'}$.

Here, τ is a pre-specific constant to denote the suitable threshold.

3 Experiments

In this section, we implement the above "two-phase" procedure and perform the experiments on three indices: Hang Seng Index (HSI), NASDAQ and FTSE 100 index (FTSE). The data are selected from the daily closing prices of the indices from September 1st to December 31th, 2003 (three months' data). The beginning four-fifth data are used for training and the rest one-fifth data are used in the one-step ahead prediction. The experimental performance is evaluated by the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE), which are frequently used as the statistical metrics.

In the experiments, the input pattern is constructed as a four-day's pattern: $\mathbf{x}_t = (y_{t-4}, y_{t-3}, y_{t-2}, y_{t-1})$. This is based on the assumption that (non)linear relationship occurs in sequential five days' prices. A commonly used function, the Radial Basis Function $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2)$, is selected as the kernel function. The margin for time t is set as $\varepsilon_t^u = \varepsilon_t^d = 0.5\rho(\mathbf{x}_t)$, where $\rho(\mathbf{x}_t)$ is the standard deviation of the input pattern at day t as justified in [9]. The parameter pair (C, γ) is set to $(4, 1)$ for HSI, $(2^5, 2^{-6})$ for NASDAQ, and $(2, 1)$ for FTSE, which are tuned by the cross-validation method. In the first phase, we construct the approximation function $f(\mathbf{x}_t)$ by performing the SVR algorithm on the normalized training data using the above settings.

After obtaining the approximation function, we observe that some training data points actually differ largely from the predictive values. We therefore in the second phase, update the corresponding ε_i^u and ε_i^d based on the proposed algorithm. The parameter τ is set to 2 for all three indices. Hence, we can deflate the influence of those differing points. A reason of τ being not so large is that the outliers still contain some useful information for constructing the approximation function and thus we cannot completely ignore them. We report the results in Table 1. The results indicate in the second phase, the prediction performance has improved on all the three indices, especially we obtain 3.45% and 5.41% improvement on the FTSE index for the RMSE and MAE criterion, respectively.

We also plot the results of NASDAQ in Fig. 2. The result of Phase I is illustrated in Fig. 2(a), while that of Phase II is in Fig. 2(b). If comparing these two figures, one can find that the approximation function (the solid line) in Fig. 2(b) is smoother than that in Fig. 2(a). Especially, the highlighted point A is a peak in Fig 2(a), but it is lowered and smoothed in Fig. 2(b). The other highlighted point B is a valley in Fig. 2(a), but it is now lifted in Fig. 2(b). This demonstrates that enlarging the margin width to the outliers can reduce their negative impact.

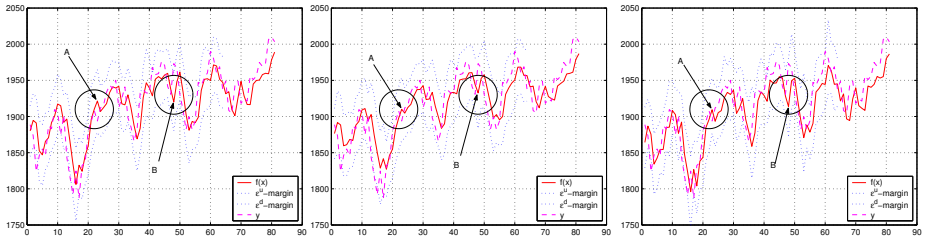
In addition, in some situations, one may prefer to predict the stock market conservatively, i.e, he would intend to under-predict the boost of stock prices for avoiding the down side risk. To meet this objective, we adopt an asymmetrical margin setting. Concretely, we pick out the corresponding up side support vectors and update their up margin and down margin by $\varepsilon_t^u = 3.8\rho(\mathbf{x}_t)$ and $\varepsilon_t^d = 0.2\rho(\mathbf{x}_t)$, respectively. Here, we use this relatively extreme setting to demonstrate the change and the difference. The graphic result is in Fig. 2(c). It can be observed that the peak A is still lowered, but the valley B is not lifted. Overall, the approximation function maintains the lower predicative values but decreases the higher predictive values, which would be highly valuable in the stock market prediction.

Table 1. Experiment Results

Dataset	Phase	RMSE	MAE	Phase	RMSE	MAE
HSI	I	140.36	116.38	II	**140.28**	**116.26**
NASDAQ	I	24.49	20.36	II	**22.78**	**19.49**
FTSE	I	59.97	44.74	II	**57.90**	**42.32**

4 Conclusion

In the paper, a novel "two-phase" SVR training procedure is proposed to detect and deflate the influence of outliers. This idea motivates from the phenomenon that enlarging the adaptive margin width in the general ε-insensitive loss function will reduce the effect of the corresponding data points. The experimental results on three indices indicate that this "two-phase" method has improvement on the prediction.



(a) (b) (c)

Fig. 2. A demonstration of the experimental results in NASDAQ. (a) is the result of Phase I. (b) is the result of Phase II with an enlarged symmetrical margin setting for the outliers detection and reduction. (c) is the result of Phase II with an enlarged asymmetrical margin setting to avoid the down side risk. The solid line is the result of the approximation function. The dashed line is the original time series. The dotted lines correspond to up margin and down margin and they are shifted away from their original places by 30, respectively, in order to make the result clear.

Acknowledgement

The work described in this paper was fully supported by two grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK4182/03E and Project No. CUHK4351/02).

References

1. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.
2. L. Cao. Support Vector Machines Experts for Time Series Forecasting. *Neurocompt.*, 51:321–339, 2003.
3. C.-C. Chang and C.-J. Lin. LIBSVM: a Library for Support Vector Machines, 2004.
4. C.-C. Chuang, S.-F. Su, J.-T. Jeng, and C.-C. Hsiao. Robust support vector regression networks for function approximation with outliers. *IEEE Transactions on Neural Networks*, 13:1322 – 1330, 2002.
5. S. Mukherjee, E. Osuna, and F. Girosi. Nonlinear Prediction of Chaotic Time Series Using Support Vector Machines. In J. Principe, L. Giles, N. Morgan, and E. Wilson, editors, *IEEE Workshop on Neural Networks for Signal Processing VII*, pages 511–519. IEEE Press, 1997.
6. J. A. K. Suykens, J. De Brabanter, L. Lukas, and J. Vandewalle. Weighted Least Squares Support Vector Machines: Robustness and Sparse Approximation. *Neurocompt.*, 2001.
7. V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.
8. V. N. Vapnik, S. Golowich, and A. Smola. Support Vector Method for Function Approximation, Regression Estimation and Signal Processing. In M. Mozer, M. Jordan, and T. Petshe, editors, *NIPS*, volume 9, pages 281–287, Cambridge, MA, 1997. MIT Press.
9. H. Yang, L. Chan, and I. King. Support Vector Machine Regression for Volatile Stock Market Prediction. *IDEAL 2002*, volume 2412 of *LNCS*, pages 391–396. Springer, 2002.
10. H. Yang, I. King, L. Chan, and K. Huang. Financial Time Series Prediction Using Non-fixed and Asymmetrical Margin Setting with Momentum in Support Vector Regression. *Neural Information Processing: Research and Development*, pages 334-350. Springer-Verlag, 2004.