A New Method for Crude Oil Price Forecasting Based on Support Vector Machines

Wen Xie, Lean Yu, Shanying Xu, and Shouyang Wang

Institute of Systems Science, Academy of Mathematics and Systems Sciences,
Chinese Academy of Sciences, Beijing 100080, China
{xiewen, yulean, xsy, sywang}@amss.ac.cn

Abstract. This paper proposes a new method for crude oil price forecasting based on support vector machine (SVM). The procedure of developing a support vector machine model for time series forecasting involves data sampling, sample preprocessing, training & learning and out-of-sample forecasting. To evaluate the forecasting ability of SVM, we compare its performance with those of ARIMA and BPNN. The experiment results show that SVM outperforms the other two methods and is a fairly good candidate for the crude oil price prediction.

1 Introduction

Oil plays an increasingly significant role in the world economy since nearly two-thirds of the World's energy consumption comes from crude oil and natural gas. Sharp oil price movements are likely to disturb aggregate economic activity, especially since Jan 2004, global oil price has been rising rapidly and brings dramatic uncertainty for the global economy. Hence, volatile oil prices are of considerable interest to many researchers and institutions. The crude oil price is basically determined by its supply and demand, but more strongly influenced by many irregular past/present/future events like weather, stock levels, GDP growth, political aspects and so on. These facts lead to a strongly fluctuating and non-linear market and the fundamental mechanism governing the complex dynamic is not understood. As Epaminondas et at. [1] reported, the oil market is the most volatile of all the markets except Nasdaq and shows strong evidence of chaos. Therefore, oil price prediction is a very important topic, albeit an extremely hard one due to its intrinsic difficulty and practical applications.

When it comes to crude oil price forecasting, most of the literatures focus only on oil price volatility analysis [1] and oil price determination within the supply and demand framework [2]. There is very limited research on oil price forecasting, including quantitative and qualitative methods. Among the quantitative methods, Huntington [3] used a sophisticated econometric model to forecast crude oil prices in the 1980s. Abramson and Finizza [4] utilized a probabilistic model for predicting oil prices and Barone-Adesi et al. [5] suggested a semi-parametric approach for forecasting oil price. Regarding the qualitative methods, Nelson et al. [6] used the Delphi method to forecast oil prices for the California Energy Commission. However, the above meth-

V.N. Alexandrov et al. (Eds.): ICCS 2006, Part IV, LNCS 3994, pp. 444 – 451, 2006.

ods show poor performance and can't meet practical needs in forecasting crude oil prices. Very recently, Wang et al. [7] proposed a new integrated methodology-TEI@I methodology and showed a good performance in crude oil price forecasting with back-propagation neural network (BPNN) as the integrated technique. BPNN, a class of the most popular neural network model, can in principle model nonlinear relations but they do not lead to one global or unique solution due to differences in their initial weight set. Another drawback is that BPNN is susceptible to over-fitting problems. Consequently, it is of necessity to develop new individual methods for forecasting oil prices which can be used for further integration into other methodologies like TEI@I.

Recently, support vector machine, a novel neural network algorithm, was developed by Vapnik and his colleagues [8]. Established on the unique theory of the structural risk minimization principle to estimate a function by minimizing an upper bound of the generalization error, SVM is resistant to the over-fitting problem and can model nonlinear relations in an efficient and stable way. Furthermore, SVM is trained as a convex optimization problem resulting in a global solution that in many cases yields unique solutions. Originally, SVMs have been developed for classification tasks [9]. With the introduction of Vapnik's ε-insensitive loss function, SVMs have been extended to solve nonlinear regression and time series prediction problems, and they exhibit excellent performance [10, 11].

The goal of this paper is to propose a new method based on SVM for the task of crude oil price time series prediction. In addition, this paper examines the feasibility of applying SVM in crude oil price forecasting through the contrast with ARIMA and BPNN models. The rest of the paper is organized as follows. Section 2 describes a new SVM-based method for crude oil price prediction. To evaluate the SVM, an empirical study and its comparable results with ARIMA and BPNN are presented in section 3. Some concluding remarks are made in section 4.

2 A New SVM-Based Crude Oil Forecasting Method

In this section, a new SVM-based method for time series forecasting and its application in crude oil price prediction are presented. We first introduce a basic theory of the support vector machine model, and then present the new SVM-based method for time series forecasting.

2.1 Theory of SVM

SVMs have originally been used for classification purposes but their principles can be extended to the task of regression and time series prediction as well. In this paper, we only focus on support vector regression (SVR) for the task of time series prediction. An excellent general introduction to SVMs including support vector classification (SVC) and support vector regression (SVR) can be seen in References [8] for more details.

SVMs are linear learning machines which means that a linear function is always used to solve the regression problem. When dealing with nonlinear regression, SVMs map the data x into a high-dimensional feature space via a nonlinear mapping φ and make linear regression in this space.

$$f(x) = (\omega \cdot \varphi(x)) + b \tag{1}$$

where b is a threshold. In linear cases, $\varphi(x)$ is just x and $f(x_i)$ becomes a linear function. Thus, linear regression in a high dimensional space corresponds to nonlinear regression in the low dimensional input space. Since $\varphi(x)$ is fixed, we determine ω from the data by minimizing the sum of the empirical risk $R_{emp}[f]$ and a complexity term $\|\omega\|^2$, which enforces flatness in feature space.

$$R_{reg}[f] = R_{emp}[f] + \lambda\|\omega\|^2 = \sum_{i=1}^{l} \psi(f(x_i) - y_i) + \lambda\|\omega\|^2 \tag{2}$$

where l denotes the sample size, $\psi(\cdot)$ is a cost function and λ is a regularization constant. For the goal of regression and time series prediction, Vapnik's ε -insensitive loss function is adopted in this paper.

$$\psi(f(x) - y) = \begin{cases} |f(x) - y| - \varepsilon & \text{for } |f(x) - y| \geq \varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

For a large set of cost functions, Eq. (2) can be minimized by solving a quadratic programming problem by applying Lagrangian theory, which is uniquely solvable under Karush-Kuhn-Tucker conditions. It can be shown that we are able to rewrite the whole problem in terms of dot products in the low dimensional input space.

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*)(\varphi(x_i) \cdot \varphi(x)) + b = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) K(x_i, x) + b \tag{4}$$

In Eq. (4), $K(\cdot)$ is the so-called kernel function which simplifies the use of a mapping. Representing the mapping by simply using a kernel is called the kernel trick and the problem is reduced to finding kernels that identify families of regression formulas. Any symmetric kernel function $K(\cdot)$ satisfying Mercer's condition corresponds to a dot product in some feature space.

2.2 SVM-Based Forecasting

The procedure of developing a support vector machine for time series forecasting is illustrated in Fig. 1. As can be seen from Fig. 1, the flow chart can be divided into four phases. The first phase is data sampling. To develop a SVM model for a forecasting scenario, training, validating and testing data need to be collected. However, the data collected from various sources must be selected according to the corresponding criteria. The second phase is sample preprocessing. It includes two steps: data normalization, data division. In any model development process, familiarity with the available data is of the utmost importance. SVM models are no exception, and data normalization can have a significant effect on model performance. After that, data collected should be split into two sub-sets: in-sample data and out-of-sample data which are used for model development and model evaluation respectively. The third phase is SVM training and learning. This phase includes three main tasks: determination of SVM architecture, sample training and sample validation. It is the core process

of SVM model. In this phase, we shall determine the time-delay τ, embedding dimension d, ε, regularization constant λ and the choice of the kernel function. The final phase is out-of-sample forecasting. When the former phases are complete, the SVM can be used as a forecaster or predictor for out-of-sample forecasting of time series.

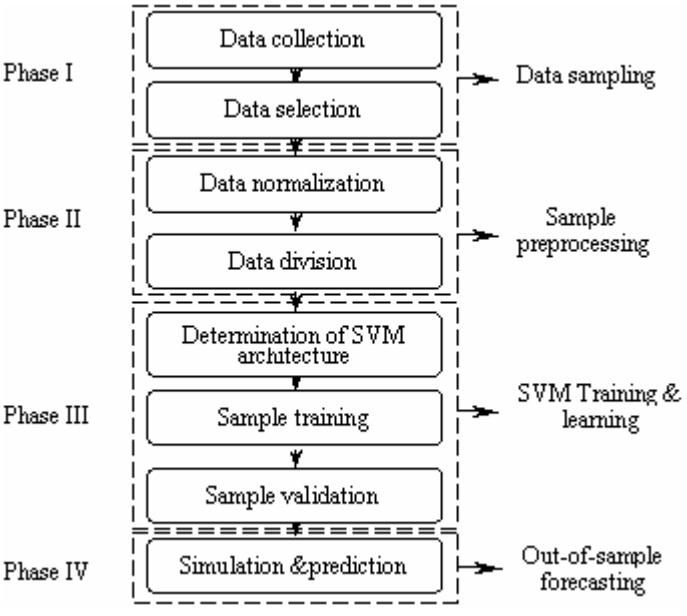

Fig. 1. A flow chart of SVM-based forecasting system

Following the above procedure, SVM-based crude oil price prediction involves four steps:

(a) Data sampling. A variety of data can be collected for this research, such as WTI, NYMEX. Data collected can be categorized into different time scales: daily, weekly and monthly. For daily data, there are various inconsistencies and missing points for the market has been closed or halted due to weekends or unexpected events. As a result, weekly data and monthly data should be adopted as alternatives.

(b) Data preprocessing. The collected oil price data may need to be transformed into certain appropriate range for the network learning by logarithm transformation, difference or other methods. Then the data should be divided into in-sample data and out-of-sample data.

(c) Training and learning. The SVM architecture and parameters are determined in this step by the training results. There are no criteria in deciding the parameters other than a trial-and-error basis. In this investigation, the RBF kernel is used because the RBF kernel tends to give good performance under general smoothness assumptions. Consequently, it is especially useful if no additional knowledge of the data is available. Finally, a satisfactory SVM-based model for oil price forecasting is reached.

(d) Future price forecasting.

3 An Empirical Study

In this section, we first describe the data, and then define some evaluation criteria for prediction purposes. Finally, the empirical results and explanations are presented.

3.1 Data

The crude oil price data used in this study are monthly spot prices of West Texas Intermediate (WTI) crude oil from January 1970 to December 2003 with a total of n = 408 observations, as illustrated in Fig. 2. Since SVM are resistant to the noise due to the use of a nonlinear kernel and an ε-insensitive band, no normalization is used in this investigation for simplicity. We take the monthly data from January 1970 to December 1999 as the in- sample data (including 60 validation data) sets with 360 observations for training and validation purposes and the remainder as the out-of-sample data sets with 48 observations for testing purposes. For space reasons, the original data are not listed here.

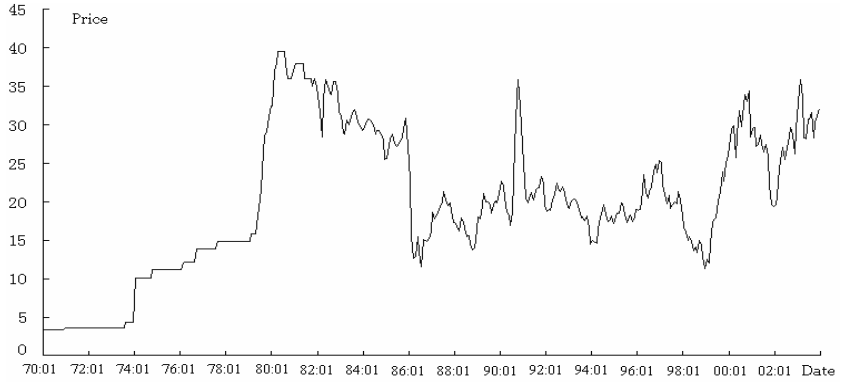

Fig. 2. The monthly oil price for the period 1970-2003

3.2 Evaluation Criteria

In order to evaluate the prediction performance, it is necessary to introduce some forecasting evaluation criteria. In this study, two main evaluation criteria, root mean square error (*RMSE*) and direction statistics (D_{stat}), are introduced. The *RMSE* is calculated as

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2} \tag{5}$$

where y_t represents the actual value, $\hat{y}_t$ is predicted values, and N is the number of testing data sets. Apparently, the indicator RMSE describes the estimates' deviation from the real values.

In the oil price forecasting, a change in trend is more important than precision level of goodness-of-fit from the viewpoint of practical applications. Trading driven by a

certain forecast with a small forecast error may not be as profitable as trading guided by an accurate prediction of the direction of movement. As a result, we introduce directional change statistics, D_{stat}. Its computational equation can be expressed as

$$D_{stat} = \frac{1}{N}\sum_{t=1}^{N} a_t \tag{6}$$

where $a_t = 1$ if $(y_{t+1} - y_t)(\hat{y}_{t+1} - y_t) \geq 0$, and $a_t = 0$ otherwise.

3.3 Results and Analysis

Each of the forecasting method described in the last section is estimated and validated by in-sample data. The model estimation selection process is then followed by an empirical evaluation which is based on the out-sample data.

The results of an augmented Dickey-Fuller (ADF) test show that the time series in level follows a unit root process. In order to utilize the ARIMA model, a first difference is necessary. Thus, ARIMA(1,1,0) is identified. The time delay τ, embedding dimension d and the prediction horizon, decided by try-and-error criteria, are respectively 4,4,1 for both BPNN and SVM. The best experiment result of each method is presented in Table 1 in which a comparison among them is performed.

Table 1. Crude oil price forecast results (Jan. 2000 - Dec. 2003)

Methods	Criteria	Full period	Sub-period I (2000)	Sub-period II (2001)	Sub-period III (2002)	Sub-period IV (2003)
ARIMA	*RMSE*	2.3392	3.0032	1.7495	1.9037	2.4868
	D_{stat} (%)	54.17	41.67	50.00	58.33	66.67
BPNN	*RMSE*	2.2746	2.9108	1.8253	1.8534	2.3843
	D_{stat} (%)	62.50	50.00	58.33	66.67	75.00
SVM	*RMSE*	2.1921	2.6490	1.8458	1.8210	2.3411
	D_{stat} (%)	70.83	83.33	50.00	58.33	91.67

Table 1 shows the detailed results of the simulated experiment via the three methods. It can be seen that the SVM method outperforms the ARIMA and BPNN models in terms of both *RMSE* and $D_{stat.}$ Focusing on the *RMSE* indicators, the values of SVM model are explicitly lower than those of ARIMA and BPNN, except in the second sub-period. From the practical application viewpoint, indicator D_{stat} is more important than indicator *RMSE*. This is because the former can reflect the trend of movements of the oil price and can help traders to hedge their risk and to make good trading decisions in advance. Concerning the SVM-based method, it can be seen in the table that although the D_{stat} value of the sub-periods II and III are some lower than BPNN, the values of D_{stat} are all above 50% and generally higher than those of the other two models, indicating that the SVM method has stronger prediction ability than the other individual models.

In addition, we observe from Table 1 that a smaller *RMSE* does not necessarily mean a higher D_{stat}. For example, for the test case of the SVM, the *RMSE* for 2001 is explicitly lower than that for 2000, while the D_{stat} for 2000 is larger than that for 2001 which implies that the D_{stat} criterion is not identical to the *RMSE* in different time series forecasting.

We can therefore conclude from the results of Table 1 and the above analysis that the proposed SVM approach performs the best among the three methods with 2.1921 and 70.83% for RMSE and D_{stat} respectively, while the ARIMA models show the worst performance.

The main reasons for the above conclusions are as follows. As Epaminondas et at. [1] reported, the crude oil market is one of the most volatile market in the world and shows strong evidence of chaos. ARIMA, typical linear models which capture time series' linear characteristics, shows insufficient to describe the nonlinear dynamics. Hence, ARIMA models perform worst among the three methods.

Both SVM and BPNN can in principle describe the nonlinear dynamics of crude oil price. Established on the unique theory of the structural risk minimization principle to estimate a function by minimizing an upper bound of the generalization error, SVM is resistant to the over-fitting problem and can model nonlinear relations in an efficient and stable way. Furthermore, SVM is trained as a convex optimization problem resulting in a global solution that in many cases yields unique solutions. Compared with the SVM's merits above, BPNN tends to suffer from over-fitting problem and does not lead to one global or unique solution owing to differences in their initial weights. Therefore, SVM generally outperforms BPNN. But, as shown in Table 1, BPNN may outperform SVM in some sub-periods. There may be two reasons: 1) the data in the sub-periods may be more suited to the BPNN's learning algorithms; 2) the chosen BPNN with the best performance outperforms SVM by chance with its initial random weight. Generally speaking, SVM outperforms ARIMA and BPNN and is more capable of oil price time series forecasting.

4 Conclusions

It has been shown in the literatures that support vector machines can perform very well on time series forecasting. The largest benefit of SVM is the fact that a global solution can be attained. In addition, due to the specific optimization procedure it is assured that over-training is avoided and the SVM solution is general.

In this paper, we propose a new method for predicting crude oil price time series based on support vector machines. There exist four phases when developing a SVM for time series forecasting: data sampling, sample preprocessing, training & learning and out-of-sample forecasting. An empirical study, in which we compare SVM's performance with those of autoregressive integrated moving average models and back-propagation neural networks, is put underway to verify the effectiveness of the SVM-based method. The results show that SVM is superior to the other individual forecasting methods in monthly oil price prediction. The prediction can be improved if irregular influences are taken into consideration in the framework of TEI@I [7], which is undoubtedly a very interesting and meaningful topic for our future study.

References

1. Panas, E., Ninni, V.: Are oil markets chaotic? A non-linear dynamic analysis. Energy Economics 22 (2000) 549-568
2. Hagen, R.: How is the international price of a particular crude determined? OPEC Review 18 (1994) 145-158
3. Huntington, H.G.: Oil price forecasting in the 1980s: what went wrong? The Energy Journal 15 (1994) 1-22
4. Abramson, B., Finizza, A.: Probabilistic forecasts from probabilistic models: a case study in the oil market. International Journal of Forecasting 11(1995) 63-72
5. Barone-Adesi, G., Bourgoin, F., Giannopoulos, K.: Don't look back. Risk August 8 (1998) 100-103
6. Nelson, Y., S. Stoner, G. Gemis, H.D. Nix: Results of Delphi VIII survey of oil price forecasts. Energy Report, California Energy Commission (1994)
7. Wang, S.Y., L.A. Yu, K.K.Lai: Crude oil price forecasting with TEI@I methodology. Journal of Systems Science and Complexity 18 (2005) 145-166
8. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, New York (1995)
9. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2 (1998) 121-167
10. Huang, W., Nakamori, Y., Wang, S.Y.: Forecasting stock market movement direction with support vector machine. Computers & Operations Research 32 (2005) 2513-2522
11. Muller, K.R., Smola, J.A., Scholkopf, B.: Prediction time series with support vector machines. Proceedings of International Conference on Artificial Neural Networks, Lausanne (1997) 999-1004