Intelligent Prediction of Crude Oil Price Using Support Vector Machines

Adnan Khashman (*Senior Member, SMIEEE*)
The Intelligent Systems Research Group (ISRG)
Near East University
Lefkosa, Mersin 10, Turkey
E-mail: khashman@ieee.org

Nnamdi I. Nwulu (*Student Member , MIEEE*)
Dept. of Electrical & Electronic Engineering
Near East University
Lefkosa, Mersin 10, Turkey
E-mail: ninwulu@ieee.org

Abstract— **The price of crude oil is tied to major economic activities in all nations of the world, as a change in the price of crude oil invariably affects the cost of other goods and services. This has made the prediction of crude oil price a top priority for researchers and scientists alike. In this paper we present an intelligent system that predicts the price of crude oil. This system is based on Support Vector Machines. Support Vector Machines are supervised learners founded upon the principle of statistical learning theory. Our system utilized as its input key economic indicators which affect the price of crude oil and has as its output the price of crude oil. Data for our system was obtained from the West Texas Intermediate (WTI) dataset spanning 24 years and experimental results obtained were very promising as it proved that support vector machines could be used with a high degree of accuracy in predicting crude oil price.**

Keywords- Crude Oil, Support Vector Machines, Statistical Learning Theory, West Texas Intermediate, Price Prediction.

I. INTRODUCTION

The advent of Globalization has led to a marked increase of the effect prices of goods and services have on one another. Top of the list of goods that have a massive effect on other goods or services is crude oil. Crude oil is arguably the most important commodity traded around the world today. Virtually every sector of the global economy is dependent on crude oil; hence any increase or decrease in the price of crude oil has a ripple effect on the global economy. In view of the importance of crude oil there has been a lot of research and attempts to predict the price of crude oil and her allied products. This is by no means a mean feat as the price of crude oil is non-linear in nature and the task is prone to many difficulties. Tools used by researchers to predict the price of crude oil can be broadly classified into two groups: Using soft computing tools or using econometric tools. In [1] we have an example of the use of soft computing tools particularly neural networks in oil price prediction. The empirical mode decomposition (EMD) technique was used. The EMD uses the Hilbert–Huang transform (HHT) and decomposes a time series to intrinsic mode functions (IMFs). A Feed forward neural network is then used to model the decomposed IMF's and residual components. The results of this neural network are then fed to an adaptive linear neural network (ALNN) which serves as an integrator. The final model is used to forecast both Brent and West Texas Intermediate (WTI) crude oil prices. The correct prediction rate in this work when using the neural network was 69%. Recently in [2] the authors developed an adaptive neurofuzzy interference system that predicts one day ahead whether the price of crude oil (WTI) is going to rise or fall. This hybrid system uses the hybrid learning algorithm and historical data from January 5th,2004 to April 30th, 2007 for training while testing was done throughout the month of May 2007 and also form May to June 2008. The correct prediction rate in this recent work was reported as 66.67%.

Support Vector Machines (SVM's) have recently been applied in fields like medicine: breast cancer diagnosis [3], Prediction: wind speed prediction [4], air pollutants level prediction [5] and even the difficult task of oil price prediction [6], [7]. In [6] the authors obtain daily West Texas Intermediate (WTI) oil prices from January 2, 2002 till October, 2008. The data is fed to a Slantlet algorithm which extracts various features as input to the hybrid system consisting of SVM's and Auto Regressive Moving Average (ARMA). In [7] the data used is monthly WTI prices from January 1970 to December 2003(a total of 408 instances) and three systems are designed to predict oil price (Neural Networks, SVM and Auto Regressive Integrated Moving Average: ARIMA). However most of the works on crude oil prediction using support vector machines suffer from a disparity in the training to testing ratio. For instance in [7] the training to testing ratio is an unbalanced 88.2%:11.8%. In [6] the training to testing ratio is 60%:40%. A more appropriate ratio would be closer to 50%:50%. This ensures that the system is not biased. Having reviewed prior works we used training to testing ratio of 50%:50% in our work. Furthermore we used weekly West Texas Intermediate (WTI) spot prices available online on the Energy Information Administration (EIA) website [8].The reason for using weekly as against daily spot prices is to minimize incomplete information due to public holidays or weekends(days when oil is not traded). The paper is organized as follows: in section two we briefly review support vector machines. In section three the oil price dataset is introduced and our novel method to prepare indicators and features from the dataset; to be used as input information for the SVM prediction model. In section four, we describe the SVM model used in this work. In section five, the

experimental results of training and testing the SVM model are discussed. Finally, in section six the paper is concluded and areas of improvement for future work are suggested.

II. Support Vector Machines (SVM)

Support Vector Machines (SVM) is an emerging tool in the soft computing field. SVM's are supervised learners founded upon the principle of statistical learning. Unlike Neural Networks and other supervised learning tools, Support Vectors have the advantage of reducing the problems of over-fitting or local minima. This is because learning in SVM is based on the structural risk minimization principle whereas in neural networks learning is based on the empirical risk minimization principle [9]. SVM works by nonlinearly mapping the inner product of a feature space to the original space with the aid a kernel. When training in SVM, the solution of SVM is unique globally, and it is only dependent on a small subset of training data points which are referred to as support vectors. SVM is capable of learning in high-dimensional spaces with a small number of training examples and has high generalization ability. There are four basic kernel types presently in use with SVM. They are the Linear Kernel, Polynomial Kernel, Radial Basis Function Kernel (RBF) and the Sigmoid Kernel. We use the RBF Kernel in this work and make use of the LIBSVM package [10] for SVM learning. The equation for the RBF kernel is given by:

$$K(x,y) = \exp(-\gamma \| x-y\|^2), \gamma > 0 \quad (1)$$

The major reason why we use the RBF kernel is because it has fewer numerical difficulties, possesses less hyper-parameters than other kernels and its ability to handle cases when the relationship between class labels and attributes is highly non-linear as in the case of Oil price prediction. To control generalization capability of SVM, the RBF kernel has two parameters: Gamma (γ) and C (cost parameter of the error term). Both C and γ should be greater than zero. The basic learning procedure using SVM is outlined in the following steps:

- Pre-process your data by scaling or normalization (we scaled all feature to values between 0 - 1)
- Consider a suitable kernel (either RBF, sigmoid, polynomial or linear)
- Obtain the best parameters by cross validation (best C and γ)
- Use the obtained best parameters (C and γ) to train the data set
- Test the trained data set on the testing dataset

In the next section we describe the crude oil price prediction dataset and our method of preparing features used in this work.

III. Dataset And Information Representation

The dataset we used during the development of the prediction system is available online by the American Energy Information Administration [8]. This dataset, which comprises weekly spot prices of West Texas Intermediate (WTI) crude oil, was from January 03 1986 to Dec 25 2009 (i.e. a total of 1252 observations). SVM's performance and efficiency is very dependent on its input features. Part of the novelty in this work is the formulation of input training factors or features from the oil prices dataset, in other words the input data coding. We propose the following input features:

1. **Years**: Ranging from 1 to 24. The year 1986 corresponds to 1, year 1987 to 2, and so on until the year 2009 which is represented with value 24.
2. **Seasonal Demand**: There is seasonality in demand for oil, where the demand in colder months is usually higher than in warmer months. We, therefore, assume two values: '1' to represent the warmer months spanning April to September, and '2' to represent the colder months spanning October to March.
3. **Average price of previous week**: The average price of the previous week is an input for the present week. This attribute is aimed at providing continuality of oil prices throughout the weeks.
4. **Total number of weeks**: Each week throughout the entire dataset is represented numerically by values from 1 to 1252.
5. **Yearly number of weeks**: Each week within each year in the dataset is numerically represented by number 1 to 53. Usually a year has 52 weeks, however, some years have 53, and thus the maximum value for this particular attribute is assigned the value 53.
6. **World events impact factor (WEIF)**: This novel attribute is aimed at accounting for the impact of political, economic and other external events on the oil price market. With such an effect being unpredictable, we use random numerical values ranging from 0.1 to 0.5; with 0.5 being a major world event that could impact the oil prices, and 0.1 a minor world event.
7. **Global demand**: Oil consumption has increased as the years go by. This has been partly ascribed to the world's increasing demand for oil, increasing Gross Domestic Product (GDP) and population growth. China, India and other Asian countries have also contributed to the increase in demand as their industrial production increased. This attribute has numerical values ranging from 1 to 7.
8. **NYMEX future contract prices**: It has been suggested in previous works that there is a relationship between spot prices and the New York Mercantile Exchange (NYMEX) future prices [11,12,13], where future contract prices are often indications of the direction spot prices would go. Therefore, we include this attribute in our prediction model design.

The above input features will be used as the input factors to the SVM prediction model during training and testing of the model. In total there are 1252 observations spanning the 24 years as indicated earlier on. Each entry in the dataset contains one set of input features and its corresponding output (the oil price). The dataset is divided equally into 626 observations for

training the model, and 626 observations for testing the trained model.

However, prior to using these observations, all values must be scaled to numerical values between 0 and 1. The data pre-processing; i.e. Feature scaling as well as output data coding, will be explained in detail in the following section.

IV. Crude Oil Price Prediction System

The developed system consists of two key phases: a scaling or normalization phase and an SVM learning phase. The scaled features obtained from the first phase serve as input values to the second phase.

A. Data Processing Phase

During this phase two important processes take place. Firstly, the values of the eight input features are scaled to values between 0 and 1. Secondly, the output values (predicted oil price) for training and testing are coded to suitable output classes. The scaling technique which is used in this work is based on finding the maximum or the highest value within each input feature for all 1252 observations in the dataset, and dividing all the values within that same feature by the obtained maximum value. Table I shows the maximum values for the input features. In order to create suitable classes for our SVM model we had to code the output. Instead of training the SVM model with output value (weekly oil price), we defined 20 weekly average price ranges within 5 US$/barrel as the output of the SVM model thus giving rise to 20 classes for our model. This is aimed at providing a degree of flexibility to the predicted price and to improve learning of the SVM model. Table II shows the price intervals and their respective output classes. Table III shows examples of the dataset observations prior to scaling; listing the first 10 observations.

TABLE I. The Highest or Maximum Value For Each Input Feature of The 1252 Observations; These Values Are Used To Normalize/Scale The Input Data Prior To SVM Training

Input Attribute	1	2	3	4	5	6	7	8
Max.Value	24	2	142.52	53	1252	0.5	7	142.46

TABLE II. Prediction System's Output Price Interval And Corresponding Output Classes

	Weekly Price Range (US$ / barrel)	Output Classes
1.	0 – 15	1
2.	16 – 20	-1
3.	21 - 25	2
4.	26 - 30	-2
5.	31 - 35	3
6.	36 - 40	-3
7.	41 - 45	4
8.	46 – 50	-4
9.	51 – 55	5
10.	56 – 60	-5
11.	61 - 65	6
12.	66 - 70	-6
13.	71 - 75	7
14.	76 - 80	-7
15.	81 - 85	8
16.	86 - 90	-8
17.	91 - 100	9
18.	100 - 110	-9
19.	111 - 120	10
20.	>120	-10

B. SVM Learning Phase

During this In the SVM learning phase we used the C-SVM model with an RBF kernel. As stated above the choice of RBF kernel is because it has fewer numerical difficulties, possesses less hyper-parameters than other kernels and is able to handle cases when the relationship between class labels and attributes is highly non-linear as in this case of Oil price prediction. In order to search for suitable parameters (C and γ) for our RBF kernel we had to perform a parameter search using cross validation specifically the *v*-fold cross validation method. The cross-validation procedure is a technique used to avoid the over fitting problem. In *v*-fold cross-validation, we first divide the training set into *v* subsets all with equal size. Sequentially one subset is tested using the SVM classifier trained on the remaining (v – 1) subsets. Cross-validation accuracy is the percentage of data which are correctly classified.

The parameters which produce the best cross validation accuracy are saved and then used to train the SVM learner. The saved model is then used on the out of sample data (testing set). In this work v=5. The parameter search range for C was conducted from (2^{-50} – 2^{50}) while for γ it was (2^{-15} – 2^{15}). The best C obtained was 2965820 while γ was 0.00195313. These values of C and γ were then used for training the SVM learner.

V. Experimental Results

The results of implementing the oil price prediction model were obtained using a 2.2 GHz PC with 2 GB of RAM, Windows XP OS and LIBSVM v 2.9.1.

There are a total of 1252 observations which comprises the weekly WTI crude oil prices from January 03 1986 to December 25 2009; available online from the USA- EIA website [8]. We used a training-to- testing ratio of (50%:50%) where the last 626 observations (period: January 02 1998 to December 25 2009) were used for training, while the first 626 observations (period: January 03 1986 to December 26 1997) were used for testing the trained prediction neural model. The reason for using the latest observations or the second period for training rather than testing is explained as follows with the aid of Fig. 1 which shows the graph of the weekly oil prices over the period from 1986 until 2009:

TABLE III. EXAMPLES OF PRE- SCALING OBSERVATIONS NUMERICAL VALUES, SHOWING THE INPUT ATTRIBUTES AND CORRESPONDING OUTPUT (OIL PRICE) FOR THE FIRST 10 OBSERVATIONS

Observations	Date	Input Features Value								Weekly Average Price (US$/barrel)
		1	2	3	4	5	6	7	8	
1.	03/01/1986	1	2	0	1	1	0.3	1	26.12	25.78
2.	10/01/1986	1	2	25.78	2	2	0.2	1	26.08	25.99
3.	17/01/1986	1	2	25.99	3	3	0.3	1	24.57	24.57
4.	24/01/1986	1	2	24.57	4	4	0.3	1	20.31	20.31
5.	31/01/1986	1	2	20.31	5	5	0.3	1	19.83	19.69
6.	07/02/1986	1	2	19.69	6	6	0.3	1	16.62	16.72
7.	14/02/1986	1	2	16.72	7	7	0.3	1	16.31	16.25
8.	21/02/1986	1	2	16.25	8	8	0.3	1	14.4	14.39
9.	28/02/1986	1	2	14.39	9	9	0.3	1	14.29	14.25
10.	07/03/1986	1	2	14.25	10	10	0.3	1	12.36	12.27

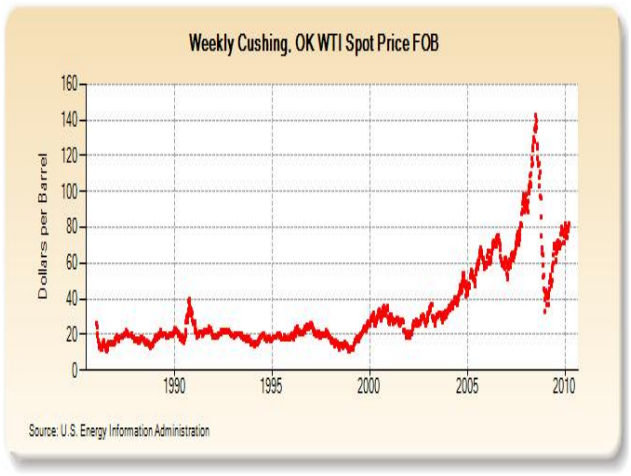

Figure 1. Graph of WTI Spot Prices (January 1986 to December 2009) [8]

- Firstly, careful observation of the graph reveals that the second period is more chaotic and the price range is more varied therefore training the SVM model with data from that period would expose it to these variations thus making it more robust when used for prediction. The first period prices can be seen as more stable and with less range variety.

- Secondly, the oil price market is an ever evolving one, it is more reasonable to train the SVM model with data from more recent periods in order to achieve a more accurate prediction of the current and future prices.

Table IV lists the final parameters of the successfully trained SVM model, and the accuracy rates. The implementation results of the trained prediction system were as follows: using the training dataset (latest 626 observations) yielded 92.7316% accuracy in prediction.

TABLE IV. FINAL PARAMETERS OF THE TRAINED SVM CRUDE OIL PRICE PREDICTION MODEL

Number of Features	8
Number of Classes	20
C parameter search range	2^{-50}–2^{50}
γ parameter search range	2^{-15}–2^{15}
C	2965820
γ	0.00195313
v	5
Training optimization Time	0.95 seconds
Training dataset prediction rate	92.7316%
Testing dataset prediction rate	69.8211%
Overall prediction rate	81.27635%
Type of SVM	C- SVM
Kernel	RBF

The testing of the trained SVM model using the testing dataset (earlier 626 observations) yielded a correct prediction rate of 69.8211%. Combining the training and testing prediction results yields an overall correct prediction rate of 81.27635%.

VI. CONCLUSION

This paper presented an SVM based prediction system with application to the difficult task of predicting oil prices. This prediction task has been addressed in few previous works that suggested using different econometric models or soft computing methods, with varying degrees of success. In this

work, we use WTI average weekly prices of crude oil over the past 24 years as the dataset for developing the prediction system. The output of the trained SVM prediction model provides the weekly average price for crude oil within 5US$/barrel accuracy.

The prediction system comprises two phases: Firstly, the data processing phase and the SVM arbitration phase. In the first phase, we apply a novel simple but efficient method of representing data information from the dataset into eight features which we believe have an effect on the oil price. The features include global demand factor and a random world event factor amongst other features. These attributes undergo scaling prior to using them as inputs to the SVM prediction system. We also presented a unique representation to the prediction output using oil prices ranges or intervals of 5 US$/barrel and created corresponding output classes. The second phase is training the SVM classifier. We used a C-SVM model with an RBF kernel. The optimal values of parameters C and γ were searched for and training this model successfully required approximately 0.95 seconds, which is considered as fast.

The obtained overall correct prediction rate of 81% is considered as successful taking into account the nonlinearity of the problem. Future work will focus on improving this prediction rate and on comparing the use of SVM to a neural network model when applied to oil price prediction.

REFERENCES

[1] L. Yu, S. Wang, and K.K. Lai, “Forecasting Crude Oil Price with an EMD-Based Neural Network Ensemble Learning Paradigm”, *Energy Economics* vol. 30, no. 5, pp. 2623–2635, 2008.

[2] A. Ghaffari and S. Zare, “A Novel Algorithm for Prediction of Crude Oil Price Variation Based on Soft Computing”, *Energy Economics* vol. 31, no. 4, pp. 531–536, 2009.

[3] M.F. Akay, "Support Vector machines combined with feature selection for breast cancer diagnosis", *Expert Systems with Applications* vol. 36, no. 2, pp.3240-3247, 2009.

[4] M.A. Mohandes, T.O. Halawani, S.Rehman and A.A.Hussain,"Support vector machines for wind speed prediction", *Renewable Energy*, vol. 29, no. 6, pp.939-947, 2004.

[5] W. Wang , C. Men and W. Lu, "Online prediction model based on support vector machines", *Neurocomputing* vol. 71, no.4,pp.550-558, 2008.

[6] K. He, C. Xie and K.K Lai, "Crude Oil Price Prediction using Slantlet Denoising Based Hybrid Method", *IEEE International Joint Conference on Computational Sciences and Optimization*, 2009.

[7] W.Xie, L.Yu, S.Xu and S.Wang. A New Method for Crude Oil Price Forecasting based on Support Vector Machines, *Lecture Notes in Computer Science*, vol.3994, pp.444-451, 2006.

[8] Energy Information Adminisntration. Weekly Cushing. OK WTI Spot Price FOB. Retrieved March 14, 2010 from the World Wide Web: “http://www.eia.doe.gov”.

[9] V. N. Vapnik. The nature of statistical learning theory.Statistics for engineering and information science. New York: Springer, 2nd ed. edition, 2000.

[10] C.C.Chang and C.-J.Lin, LIBSVM: a library for support vector machines,2001. Software available at: /http://www.csie.ntu.edu.tw/cjlin/libsvm.

[11] B.N. Huang, C.W. Yang, and M.J. Hwang, “The Dynamics of a Nonlinear Relationship Between Crude Oil Spot and Futures Prices: A Multivariate Threshold Regression Approach”, *Energy Economics*, vol. 31, no. 1, pp. 91–98, 2009.

[12] S. Dées, A. Gasteuil, R.K. Kaufmann and M. Mann,“Assessing the Factors Behind Oil Price Change”. *European Central Bank Working Paper Series*, no.855, 2008

[13] S.D. Bekiros and C.G.H. Diks,“The Relationship Between Crude Oil Spot and Futures Prices: Cointegration, Linear and Nonlinear Causality”, *Energy Economics*, vol. 30, no. 5, pp. 2673–2685, 2008.