# Oil Price Prediction Using Ensemble Machine Learning

Lubna A.Gabralla[1], Rania Jammazi[2] and Ajith Abraham[3, 4]

[1]Faculty of Computer Science & Information Technology, Sudan University of Science Technology, Khartoum, Sudan
lubnagabralla@gmail.com
[2]Faculty of Management and Economic Sciences of Sousse, El-Riadh City, Sousse University, Tunisia
jamrania2@yahoo.fr
[3]Machine Intelligence Research Labs (MIR Labs) Scientific Network for Innovation and Research Excellence, WA, USA
[4]IT4Innovations, VSB-Technical University of Ostrava, Czech Republic
ajith.abraham@ieee.org

*Abstract*— Crude oil price forecasting is a challenging task due to its complex nonlinear and chaotic behavior. During the last couple of decades, both academicians and practitioners devoted proactive knowledge to address this issue. A strand of them has focused on some key factors that may influence the crude oil price prediction accuracy. This paper extends this particular branch of recent works by considering a number of influential features as inputs to test the forecasting performance of daily WTI crude oil price covering the period 4th January 1999 through 10th October 2012. Empirical results indicate that the proposed methods are efficient and warrant further research in this field.

*Index Terms*— crude oil price prediction; hybrid models; influential features

## I. Introduction

II. Oil is one of the most important and valuable natural resources in the world economy. According to the Energy Information Administration (EIA), the world currently consumes 85.64 million barrels of crude oil daily. That is about 2 liters of oil for every single person on the planet every day [1]. Therefore, it has been sometimes called "black gold" or "life blood". Most countries heavily relied on imported crude oil in order to meet their energy needs. Oil exporting countries attempt to use oil as a weapon to perpetuate and wield political and economic power. Changes in world crude oil prices are becoming an increasing source of concern for government's economic and organizational decisions. Knowing that every economic sector in the world is dependent on crude oil; hence any increase or decrease in the price of crude oil has a ripple effect on the global economy [2]. Researchers provided some possible evidences of this effect. Jain [3] examined the linkage between global oil prices, precious metal prices (Gold, Platinum and Silver) and Indian Rupee – US Dollar exchange rate. Authors' result shows the presence of a significant relationship between the precious metals and oil. Malliaris and Malliaris [4] explored the inter–relationships between two commodity prices (gold and oil) and the euro using standard time series and neural network methodologies. Their results indicated that oil impacts gold more than gold impacts oil, while oil's effect on the euro is greater than the euro's effect on oil. Doğrul and Soytas [5] investigated the relationship between oil prices, unemployment and interest rate in Turkey. Therefore the behavior of oil prices and the factors affecting them are receiving a special international attention. Crude oil prices prediction is not an easy task because there are many factors that can influence their tendency such government interventions, political events, weather conditions, financial speculations, supply, inventories, demand, exchange rates, OPEC oil policy, GDP, financial shocks, price trends and stock market, dollar index, gold, heating oil spot price, etc. [6-8]. The main purpose of our research is to apply machine-learning approaches to forecast crude oil prices and address the following questions: Can machine learning based models predict the crude oil price accurately? Which set of features can better describe the performance? Can we achieve comparable or relatively high prediction performances by introducing time lags? The structure of this paper is as follows. Section 2 depicts the literature review followed by the oil price depending factors and the role of feature selection in Section 3. Different forecasting models are presented in Section 4. Section 5 provides the details about data preprocessing, experimental results; discussions and conclusions are provided towards the end.

## II. Related Research

With inadequate information, too many variables, and imprecise elements, the oil price system is extremely complex for modeling analytically, and its dynamics are hard to predict [9]. Forecasting oil prices brought a considerable attention by researchers who provide a proactive knowledge in identifying potential candidate forecasting models for crude oil prices. In their pioneering study, Abramson and Finizza [10-12] used Belief Networks (BNs) to forecast crude oil. More recently, they used rather a probabilistic belief network model to address the given question [13]. Morana [14] proved that Generalized Autoregressive Conditional Heteroskedastic (GARCH) models are more convenient to forecast the oil prices than a simple random walk model. The superiority of the model is deduced by means of the decomposition of the mean square forecast error. Lanza et al. [15] investigated the relationships between heavy crude oil and products price using co integration and error correction models and evaluated the predictive power of the specification in forecasting crude oil prices. However, previous studies described the behavior of oil price as non-linear and econometric and statistical systems are able to

achieve logical results in the case of linear behavior [16]. As a result, new techniques such as artificial neural networks, genetic algorithm and support vector machine have emerged to remedy to this inefficiency [17]. Alizadeh and Mafinezhad [17] applied a General Regression Neural Network (GRNN) model to forecast Brent oil price at the short term. By including seven types of features as inputs, authors claimed that their constructed model has turned out to provide good deal of precision under critical conditions. Malliaris and Malliaris [4] used a neural network to study relationships among the price behavior of gold, oil and the euro using daily data over the period started from 4th January 2000 to 31st December 2007. Their results illustrate that oil price plays a significant role in predicting gold price and Euro exchange rate. Based on the fact that ANN models often suffer from the local minima and over fitting problems [18], Adnan and Namdi [2] designed an intelligent system based on Support Vector Machines to predict the price of crude oil involving eight input factors (global demand; a random world event; among others). Empirical results show high prediction accuracy. However, previous studies have shown that hybrid models achieve more accurate results than individual traditional models [5]. Among others, Jammazi and Aloui [6] proposed wavelet based –MBPNN (multilayer back propagation neural network) model for crude oil price forecasting by modifying the neural mapping. This model combines the dynamic properties of MBPNN and Haar A Trous wavelet (HTW) decomposition. Using the combination between Genetic Algorithm and Support Vector Machine (GA-SVM) and based on RMSE (Root Mean Squared Error) criteria, Guo et al. [8] proved the superiority of the proposed one over the standard SVM to forecast daily Brent oil stock price data for the period running from 2000 to 2011.

## III. Feature Selection for Forecasting

Oil prices have remained hard to predict due to its complexity and irregularity. The complexity is mainly due to its dependence on many global and national economic factors. The magnitude of these linkages is difficult to quantify since oil prices and external factors form a complicated network structure itself promote direct/indirect and repetitive/cyclical oscillations of the given signals. With regard to the approaches mentioned above, almost all of them are inadequate to accurately design this implementing neural architecture. The main reason is attributed to the irregularity and the sudden abrupt changes in the oil price behavior that marked the last three decades [19]. Specifically, the prediction will inevitably be incomplete, as the network representation of the relationships between oil prices and the respective factors require an explicit mapping [20]. Numerous studies have focused on developing new techniques but little attention has been paid to test the predictive power of different inputs. The desired output of any proposed model (linear or nonlinear) heavily depends on the information content of the inputs [20]. Generally, economic and market factors are considered to be the main causes of surging and falling global oil prices [22]. Shin et al. [19] design a graph-based semi supervised learning approach to represent the relationship between some factors. On one hand, they used the association of the demand-related variables including the overall international, OPEC as well as Saudi oil productions. On the other hand, the supply-related variables involve the overall international, OECD as well as non-OECD countries' oil demands. Moreover, they consider other economic indicators such as the producer price indices, the US dollar exchange rates, OECD commercial stockpiles, NYMEX oil futures price and the transaction volume in the WTI (West Texas Intermediate Cushing) crude oil price. The main findings of their investigations are that input variables affect the WTI and vice-versa. They focus only on some commonly intrinsic features to avoid any unnecessary increases in the dimensionality of the input space, which degrades the prediction performance of the model. Haidar et al. [20] divided features in two groups. The first one assembles crude oil futures prices while the second one contains inter-market variables, such as S&P 500 (to represent the market performance), Dollar index (to show the strength of the USD), gold prices which is deemed to be less volatile than crude oil prices and heating oil (to highlight seasonal information presented in energy market). Authors used feed forward back propagation NN composed of three layers. The proposed forecasting tool appears to support that heating oil spot price has significant explanatory power for crude oil spot price for multiple step predictions. Using crude oil benchmark markets namely (WTI) and European Brent crude oil (Brent), He et al. [21] build an algorithm based on wavelet. The experiment results show the superiority of the proposed algorithm over the benchmark models in terms of both level and prediction accuracy. Mingming et al. [22] propose multiple adaptive wavelet recurrent neural networks to predict crude oil prices and conclude that it is possible to use commercial material instead of gold price for crude oil prediction. Azadeh et al. [23] designed a flexible algorithm based on artificial neural network (ANN) and fuzzy regression (FR) to forecast oil price using some economic indicators such as oil supply, crude oil distillation capacity, oil consumption of non-OECD, USA refinery capacity, and surplus capacity.

## IV. Predictive Models Used

### *A. Support Vector Regression*

Support Vector Machines (SVM) are supervised learning models used for classification and regression analysis. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. Support Vector Regression (SVR) is a SVM algorithm to handle non-linear prediction. SMOreg (Sequential Minimal Optimization for regression) is an iterative optimization algorithm proposed by Smola and Scholkopf [25] for using SVR regression. SMOreg uses constraints structural risk minimization as the model and has the good ability to model regression, prediction with non-linear data.

### *B. Instance Based Learning*

Instance-based learning (IBL) algorithms are derive from the nearest neighbor machine learning philosophy. IBK is an implementation of the k-nearest neighbor's algorithm [26].

The number of nearest neighbors (k) can be set manually or determined automatically. Each unseen instance is always compared with existing ones using a distance metric.

*C. K Star*

K Star (K*) is an instanced based classifier [27]. A new data instance is classified by comparing it to the stored examples in order to find the most similar ones. This approach is also called nearest neighbor classification and the main advantage of this approach is that arbitrary complex structures in the data can be captured and training and retraining this model is fast.

D. *The Ensemble method*

In this research, we employed the basic ensemble method (BEM) defined by:

$$f_{BEM} = \frac{1}{n}\sum_{i=1}^{n} f_i\,(x)$$

Where *fi (x)* is the output produced by the different models. This approach by itself can lead to improved performance, but does not take into account the fact that some networks may be more accurate than others. It has the advantage of being easy to understand and implement and is often found not to increase the expected error [29].

We first constructed the IBL and SMOReg models to obtain a very good generalization performance. Then the ensemble approach was used for 3, 4 and 5 attributes as illustrated in Figure 1.

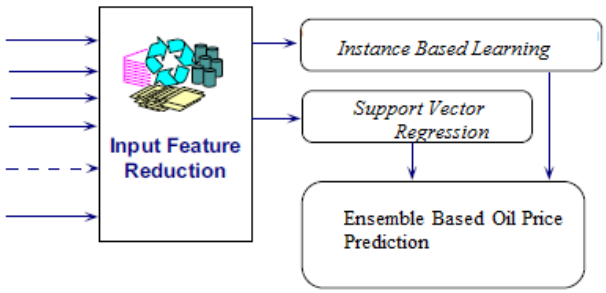


**Figure 1.** Ensemble approach combining IBL and SMOreg

## V. Experimental Results and Analysis

*A. Data set and Experimental Environment*

The daily data from 1999 to 2012 [29] to predict the West Taxes Intermediate (Output) was used. Following are the input variables:

- Cushing, OK Crude Oil Future Contract1 (Dollars per Barrel);
- Cushing, OK Crude Oil Future Contract 2(Dollars per Barrel);
- Cushing, OK Crude Oil Future Contract 3(Dollars per Barrel);
- Cushing, OK Crude Oil Future Contract 4 (Dollars per Barrel).
- Date

In the pre-processing step, the data is filtered to remove the noise and improve the quality. Attributes (Feature) selection is focused on removing those features, which do not contribute to the enhancement of the prediction. The experiments are accomplished in Version 3.6 of WEKA (Waikato Environment for Knowledge Analysis) [30].

*B. Feature Selection Methods*

We used the Best - first, Genetic algorithm based search methods for attribute selection. Best-first search is a method that does not just terminate when the performance starts to drop but keeps a list of all attribute subsets evaluated so far, sorted in order of the performance measure, so that it can revisit an earlier configuration instead. Given enough time it will explore the entire space, unless this is prevented by some kind of stopping criterion.

Genetic algorithm based search procedures are loosely based on the principal of natural selection: they evolve good feature subsets by using random perturbations of a current list of candidate subsets [30]. The above two methods helped to reduce the number of features to three, four and five attributes respectively:

*C. Influence of Data Partitioning*

We also investigated the effect of training and testing data by randomly splitting them as follows:

- 90% - 10% (A)
- 80% - 20% (B)
- 70% - 30% (C)
- 60% - 40% (D)

*D. Discussions and Analysis*

Table 1 reports the empirical results illustrating the root mean squared error (RMSE) and Correlation Coefficient (CC) with three, four and five attributes.

As illustrated in Table 1, the K Star algorithm did not perform well for all the training and testing combinations and for the three different attributes. For this reason, we excluded K Star from the simulation process. Both SMOreg and IBL exhibited the best performance in the case of three attributes with the combination A. SMOReg performed slightly well than IBL with a lower RMSE while the ensemble of them provided higher accuracy only for the case of three attributes also with combination A (RMSE of 0.326 and a CC of 0.9999). In addition, it seems that more is the percentage of training data (90% training and 10% test data) for the ensemble better is the learning. However, when the amount of training data is reduced, the ensemble requires greater number of attributes to achieve higher performance. Figures 2 - 7 illustrate the comparison among the four different algorithms for selected sample instances and different number of attribute sets.

## VI. Conclusions

In this paper, we investigated the performance of several machine-learning methods for the prediction of crude oil prices. We used three different algorithms for feature (attribute) selection. We considered IBL, KStar and SMOReg models for oil price prediction and then an ensemble model was constructed.

| Classifier | Data | RMSE | CC | RMSE | CC | RMSE | CC |
|---|---|---|---|---|---|---|---|
| | | Three attributes | | Four attributes | | Five attributes | |
| Kstar | A | 5.006 | 0.9888 | 2.084 | 0.9981 | **1.182** | 0.9996 |
| | B | **0.523** | 0.9998 | 1.386 | 0.9991 | 1.206 | 0.9996 |
| | C | 5.396 | 0.9880 | 36.818 | 0.1733 | **0.746** | -0.0136 |
| | D | 24.033 | 0.6250 | **2.374** | 0.9977 | 40.766 | 0.9996 |
| SMOReg | A | **0.381** | 0.9999 | 0.564 | 0.9998 | 0.792 | 0.9993 |
| | B | **0.467** | 0.9999 | 0.923 | 0.9995 | 1.042 | 0.9992 |
| | C | **0.578** | 0.9998 | 0.980 | 0.9994 | 1.113 | -0.0129 |
| | D | 31.610 | 0.4017 | **0.744** | 0.9997 | 41.132 | 0.9997 |
| IBL | A | **0.386** | 0.9999 | 0.451 | 0.9998 | 0.669 | 0.9996 |
| | B | **0.523** | 0.9998 | 0.718 | 0.9996 | 0.724 | 0.9996 |
| | C | **0.622** | 0.9999 | 0.742 | 0.9996 | 0.746 | 0.9997 |
| | D | 0.668 | 0.9997 | **0.660** | 0.9997 | 0.712 | 0.9998 |
| Ensemble | A | **0.326** | 0.9999 | 0.393 | 0.9999 | 0.523 | 0.9996 |
| | B | **0.419** | 0.9999 | 0.663 | 0.9997 | 0.722 | 0.9996 |
| | C | **0.528** | 0.9998 | 0.732 | 0.9996 | 0.793 | 0.7067 |
| | D | 15.823 | 0.8393 | **0.608** | 0.9997 | 20.587 | 0.9996 |

**Table 1:** Evaluation of different algorithms for oil prediction

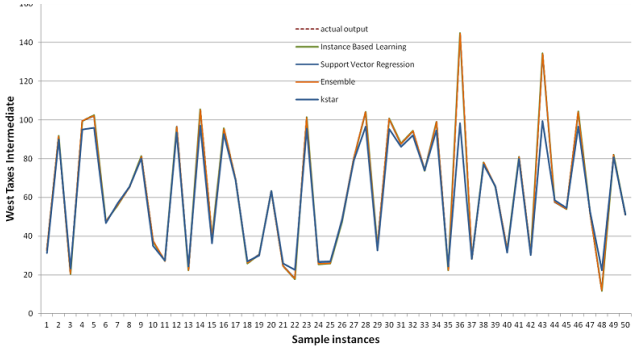


**Figure 2:** Three attributes and 90% training data

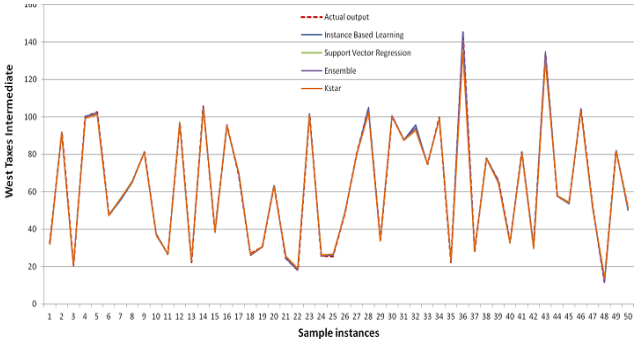


**Figure 4:** Five attributes and 90% training data

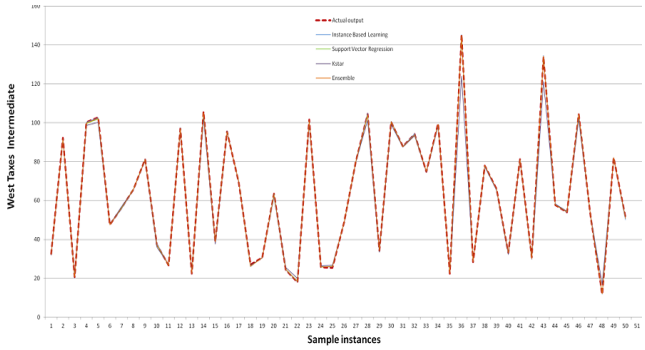


**Figure 3:** Four attributes and 90% training data

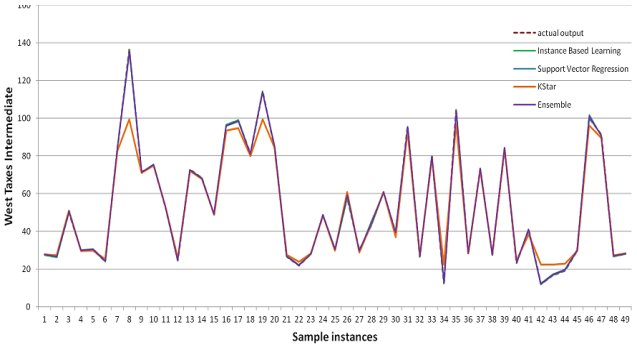


**Figure 5:** Three attributes and 60% training data

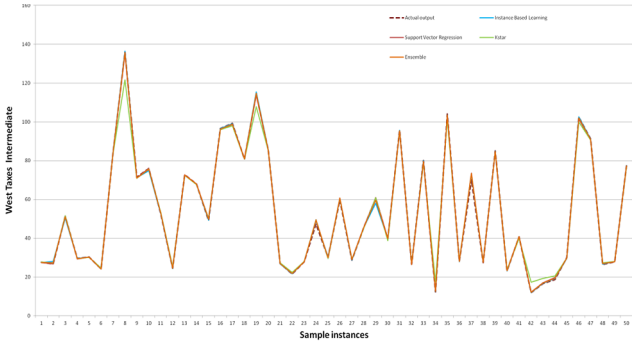


**Figure 6:** Four attributes and 60% training data

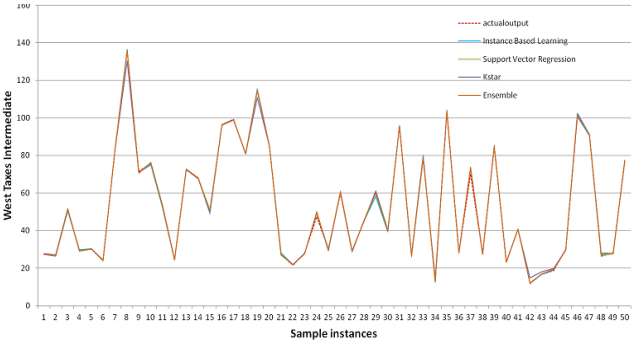


**Figure 7:** Five attributes and 60% training data

We also illustrated the effect of different subsets of training and testing data by randomly splitting them into four different groups. Empirical results illustrate that the developed ensemble method performed better than SMOReg and IBL using only three attributes. In our future research, we will investigate more hybrid models and ensemble approaches considering dynamic learning of weights.

## Acknowledgments

This work was partially supported in the framework of the IT4 Innovations Centre of Excellence project, reg. no. CZ.1.05/1.1.00/02.0070 by operational programme 'Research and Development for Innovations' funded by the Structural Funds of the European Union and state budget of the Czech Republic, EU.

## References

[1] www.suite101.com /article/oils.importance-to-the-world-economy

[2] Khashman, Adnan, and Nnamdi I. Nwulu. "Intelligent prediction of crude oil price using Support Vector Machines." Applied Machine Intelligence and Informatics (SAMI), 2011 IEEE 9th International Symposium on. IEEE, 2011.

[3] Jain, Anshul, and Sajal Ghosh. "Dynamics of global oil prices, exchange rate and precious metal prices in India." *Resources Policy* (2012).

[4] Malliaris, A. G., and Mary Malliaris. "Time series and neural networks comparison on gold, oil and the euro." Neural Networks, 2009. IJCNN 2009. International Joint Conference on. IEEE, 2009.

[5] Doğrul, H. Günsel, and Ugur Soytas. "Relationship between oil prices, interest rate, and unemployment: Evidence from an emerging market." Energy Economics 32.6 (2010): 1523-1528.

[6] Jammazi, Rania, and Chaker Aloui. "Crude oil price forecasting: Experimental evidence from wavelet decomposition and neural network modeling." Energy Economics 34.3 (2012): 828-841.

[7] Haidar, Imad, Siddhivinayak Kulkarni, and Heping Pan. "Forecasting model for crude oil prices based on artificial neural networks." Intelligent Sensors, Sensor Networks and Information Processing, 2008. ISSNIP 2008. International Conference on. IEEE, 2008.

[8] Guo, Xiaopeng, DaCheng Li, and Anhui Zhang. "Improved Support Vector Machine Oil Price Forecast Model Based on Genetic Algorithm Optimization Parameters." *AASRI Procedia* 1 (2012): 525-530.

[9] Yi, Yao, and Ni Qin. "Oil price forecasting based on self-organizing data mining." *Grey Systems and Intelligent Services, 2009. GSIS 2009. IEEE International Conference on*. IEEE, 2009.

[10] Abramson, Bruce, and Anthony Finizza. "Using belief networks to forecast oil prices." International Journal of Forecasting 7.3 (1991): 299-315.

[11] Abramson, B., and A. J. Finizza. "A belief network implementation of target capacity utilization." Proceedings of the 13th North American Conference of the International Association for Energy Economics. 1991.

[12] Abramson, B., and A. J. Finizza. "A Belief Network-Based System that Forecasts the Oil Market by Constructing Producer Behavior." Proceedings of the 15th North American Conference of the International Association for Energy Economics. 1993.

[13] Abramson, Bruce, and Anthony Finizza. "Probabilistic forecasts from probabilistic models: a case study in the oil market." International Journal of forecasting 11.1 (1995): 63-72.

[14] Morana, Claudio. "A semiparametric approach to short-term oil price forecasting." Energy Economics 23.3 (2001): 325-338.

[15] Lanza, Alessandro, Matteo Manera, and Massimo Giovannini. "Modeling and forecasting cointegrated relationships among heavy oil and product prices." Energy Economics 27.6(2005): 831-848.

[16] Liu, Jinlan, Yin Bai, and Bin Li. "A new approach to forecast crude oil price based on fuzzy neural network." Fuzzy Systems and Knowledge Discovery, 2007. FSKD 2007. Fourth International Conference on. Vol. 3. IEEE, 2007.

[17] Alizadeh, A., and Kh Mafinezhad. "Monthly Brent oil price forecasting using artificial neural networks and a crisis index." Electronics and Information Engineering (ICEIE), 2010 International Conference On. Vol. 2. IEEE, 2010.
[18] Wang, Jue, Wei Xu, Xun Zhang, Yejing Bao, Ye Pang, and Shouyang Wang. "Data Mining Methods for Crude Oil Market Analysis and Forecast." Data Mining in Public and Private Sectors: Organizational and Government Applications (2010): 184
[19] Shin, Hyunjung, Tianya Hou, Kanghee Park, Chan-Kyoo Park, and Sunghee Choi. "Prediction of Movement Direction in Crude Oil Prices Based on Semi-Supervised Learning." Decision Support Systems (2012).
[20] Haidar, Imad, Siddhivinayak Kulkarni, and Heping Pan. "Forecasting model for crude oil prices based on artificial neural networks." Intelligent Sensors, Sensor Networks and Information Processing, 2008. ISSNIP 2008. International Conference on. IEEE, 2008.
[21] He, Kaijian, Lean Yu, and Kin Keung Lai. "Crude oil price analysis and forecasting using wavelet decomposed ensemble model." Energy (2012).
[22] Mingming, Tang, and Zhang Jinliang. "A multiple adaptive wavelet recurrent neural network model to analyze crude oil prices." Journal of Economics and Business (2012).
[23] Azadeh, Ali, Mohsen Moghaddam, Mehdi Khakzad, and Vahid Ebrahimipour. "A flexible neural network-fuzzymathematical programming algorithm for improvement of oil price estimation and forecasting." Computers & Industrial Engineering 62, no. 2 (2012): 421-430.
[24] Look what you've done! Task recognition based on PC activities Saskia Koldijk Radboud University Nijmegen The Netherlands SaskiaKoldijk@student.ru.nl s061003 June 2011.
[25] Alex J. Smola and Bernhard Schölkopf, A tutorial on support vector regression, Statistics and Computing Archive, Volume 14 Issue 3, pp. 199-222, 2004.
[26] D. Aha, D. Kibler, Instance-based learning algorithms. Machine Learning. 6:37-66, 1991.
[27] John G. Cleary, Leonard E. Trigg: K*: An Instance-based Learner Using an Entropic Distance Measure, 12th International Conference on Machine Learning, 108-114, 1995.
[28] Maqsood, Imran, Muhammad Riaz Khan, and Ajith Abraham. "An ensemble of neural networks for weather forecasting." *Neural Computing & Applications* 13, no. 2 (2004): 112-122.
[29] http://www.eia.gov
[30] Witten, Ian H., and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.