Decision Support Systems 55 (2013) 348–358

Contents lists available at SciVerse ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Prediction of movement direction in crude oil prices based on semi-supervised learning

Hyunjung Shin [a,*,1], Tianya Hou [b,1], Kanghee Park [a], Chan-Kyoo Park [c], Sunghee Choi [d]

[a] Department of Industrial Engineering, Ajou University, San 5 Wonchun-dong, Yeongtong-gu, Suwon, 443-749, Republic of Korea
[b] Department of Building and Real Estate, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
[c] Department of Management, Dongguk University-Seoul Campus, 3-26 Pildong, Chung-gu, Seoul, 100-715, Republic of Korea
[d] Department of International Commerce, Keimyung University, 2800 Dalgubeoldaero, Dalseo-gu, Daegu, 704-701, Republic of Korea

ARTICLE INFO

Article history:
Received 21 May 2011
Received in revised form 21 March 2012
Accepted 4 November 2012
Available online 12 November 2012

Keywords:
Oil price prediction
Semi-supervised learning (SSL)
Technical indicators
Feature extraction (PCA/NLPCA)
Machine learning

ABSTRACT

Oil price prediction has long been an important determinant in the management of most sectors of industry across the world, and has therefore consistently required detailed research. However, existing approaches to oil price prediction have sometimes made it rather difficult to implement the complex interconnected relationship between the price of oil and other global/domestic economic factors. This has been complicated by the influence of the irregular impact caused by the economic factors that affect the oil price. Recently, a machine learning algorithm, known as semi-supervised learning (SSL) has emerged, whose strength is the ease it can bring to the network representation of entities and the explicitness of inference which is expressed through relations between different entities. Since an awareness of the network representation of complicated relations between economic factors including the oil price is natural in SSL, this method allows the effects of the impact of economic factors on the oil price to be assessed with improved accuracy. SSL has so far been exploited in dealing with the non time-series types of entity, but not for the time-series types. Therefore, the proposed study is to exploit the method of representing the network between these time-series entities, and to then employ SSL to forecast the upward and downward movement of oil prices. The proposed SSL approach will be tested using one-month-ahead monthly crude oil price predictions between January 1992 and June 2008.

1. Introduction

For many years, the price of crude oil price has been an important determinant of global or national economic performance. An increase or decrease in the oil price will have a marked economic effect on all countries in the world. Thus the state of oil prices is a consistent preoccupation of economic experts in most industries, as well as many politicians. A change in the crude oil price can lead to a transfer of income between importing and exporting countries through a shift in the terms of trade [7], since oil is the world's most actively traded commodity accounting for over 10% of total world trade [27]. For net oil-importing countries, higher oil prices can lead to a drop in real national income caused by increased input costs, along with reduced non-oil demand, higher inflation, lower investment, upward pressure on wage levels, higher unemployment, reduced tax revenues, increases in budget deficits, higher interest rates, and downward pressure on exchange rates. Net oil-exporting countries may experience those economic phenomena in the opposite way. An increase in the oil price can directly boost the real national income of net oil-exporting countries because of the higher earnings they will get from exports, which may eventually lead to greater concentration of international assets. Over the longer term, part of this gain may be later offset by losses from the lower demand for exports due to the economic recession propagated by trading partners. The bigger the crude oil price increase and the longer higher prices are sustained, the bigger the impact on the global economy.

The overall mechanism by which the oil price affects most global or national economic factors is generally well understood, and therefore forecasting the oil price has been perceived as an important research topic. One of the most commonly used approaches to oil price prediction is the statistical time-series method [20,28], which characterizes the oil price as consisting of a time trend, a seasonal factor, a cyclical element and an error term. Many techniques are available to break up a series of oil prices into these components. They include Akarca and Andrianacos' autoregressive integrated moving average (ARIMA) model [2], Lanza et al.'s error correction model (ECM) [21], and Mirmirani and Li's vector auto-regression (VAR) model [24]. Other kinds of approach assume that stochastically quantifying the relationship between the oil price and the latent economic factors may provide more relevant prediction than attempting to uncover the underlying structure of the series itself. Such methods

* Corresponding author. Tel.: +82 31 219 2417; fax: +82 31 219 1610.
E-mail addresses: shin@ajou.ac.kr (H. Shin), tianya.hou@connect.polyu.hk (T. Hou), can17@ajou.ac.kr (K. Park), parkck@dongguk.edu (C.-K. Park), choisu@kmu.ac.kr (S. Choi).
[1] Both authors contributed equally.

0167-9236/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.dss.2012.11.009

include the stochastic, semi-parametric, and wavelet-based methods. Cortazar and Schwartz implemented a stochastic model for oil futures prices [11], Morana suggested a semi-parametric statistical method for short-term forecasting based on the GARCH properties of crude oil price [25], and Yousefi et al. applied a wavelet-based technique to predict crude oil prices [39]. Other approaches using data mining or machine learning algorithms have also been applied to oil price prediction problem. Yu et al.'s ensemble learning method, which is based on artificial neural network (ANN) [40] and Xie et al.'s support vector machines (SVM) [37].

Despite such attempts, oil price prediction has remained a difficult problem due to its *complexity* and *irregularity*. The *complexity* is mainly due the complex interactions of many global and national economic factors. Such influences are certainly significant but their magnitude is difficult to quantify because the relationship between the oil price and external factors is a complicated network structure which is affected by direct/indirect and repetitive/cyclic influences. As regards the approaches mentioned above, most techniques are difficult to use in implementing the network structure. On the other hand, the *irregularity* is caused by the sudden movement of the oil price as a number of sharp price increases/decreases have occurred in the past. There have been several oil price shocks in 1973, 1978 and 2008, including drastic price collapses in 1986 and 1998 [1]. The price of oil is basically determined by balancing the amount of oil the net oil-exporting countries can supply with the demands of the net importing countries, but the irregularity is caused more by the shocks on the supply-side, which may be political disputes or sudden changes in external economic factors [3,6,12,16,17,34]. In such cases, a precise prediction of the values of the oil price will be difficult to obtain. However, a rough prediction of the upward and downward changes of the price can still be helpful for decision making. Some of the approaches mentioned above can predict a binary estimate for the changes. But the prediction will inevitably be incomplete, as a propagation pathway of the irregular impact of the source factor on the oil price is difficult to illustrate without explicit network representation on the relationship.

Most recently, a category of machine learning algorithms, known as semi-supervised learning (SSL) has emerged, the main strength of which is that it allows taking advantage of the strengths of both supervised learning and unsupervised learning[10,44]. The primary goal of supervised learning is to build accurate classifiers or regressors using labeled data. On the other hand, unsupervised learning is usually employed to discover data structure from unlabeled data. In semi-supervised learning, meaningful representation of complicatedly structured data is identified from unlabeled data, and then the decision or regression function is achieved on both labeled and the unlabeled, which is smooth with respect to the underlying geometry. SSL is regarded as a more pragmatic learning scheme since many practical domains are in such situation that there is a large supply of unlabeled data but limited labeled data which can be expensive, difficult, and time-consuming to generate. Many related researches have shown validity of SSL in a number of application domains such as spam filtering[43], document categorization [30], video surveillance [31], text classification [35], text chunking [4], gene expression data classification [5,13], and webpage classification [22], etc. In those literatures, SSL is often compared with the representative models of supervised learning, and shows its superiority over them thanks to its capability of learning from only a few labeled data utilizing a large amount of unlabeled data. There has been a whole spectrum of interesting ideas on how to learn from both labeled and unlabeled data, e.g., the expectation-maximization based approach [26], self–training [38], co-training [9], Transductive support vector machines [18], and the graph-based approaches such as graph mincuts [8], harmonic approach [45], and local and global consistency [41], etc. Among several types of SSL algorithms, a graph-based SSL is employed in our study [32,33]. In graph-based SSL, the entities are connected via the similarities between them, and prediction about an entity is made by assessing the propagated influence of its neighboring entities through the connections that exist within them.

In this paper, we propose a graph-based SSL approach for predicting upward and downward changes in a series of oil prices. The representation of complicated relations between entities is natural in the SSL learning framework and the propagation of a change in an entity is explicitly elucidated via network structure. By treating the economic factors, including the oil price, as entities of the network, these features of SSL will contribute to resolving the complexity and irregularity of the oil price prediction problem. SSL has been exploited to some extent for assessing the non time-series types of entity, but not for time-series types. Therefore, the intention here is to exploit this method of representing the relationship between time-series type entities and to then employ SSL to forecast the upward and downward movement of oil prices. The proposed SSL approach will be applied to the crude oil price prediction of West Texas Intermediate (WTI) from January 1992 to June 2008, and will be validated through comparison with an autoregression model, a logistic regression model, an ANN model and an SVM model.

The rest of this paper is organized as follows. Section 2 briefly introduces the SSL algorithm. Section 3 presents the proposed SSL model for time series prediction. Section 4 provides the experimental results as evaluated with regard to the WTI crude oil prices and compares the proposed model with five other representative models. Finally, in Section 5, conclusions will be drawn.

2. Semi-supervised learning

In graph-based SSL algorithm, a data point (or entity) $x_i \in R^M (i = 1,\ldots, n)$ is represented as a node i in a graph (or network), and the relationship between data points is represented by an edge where the connection strength from each node j to each other node i is encoded as w_{ij} of a similarity matrix W [42]. Fig. 1 presents a graphical representation of SSL.

A weight w_{ij} can take a binary value (*0* or *1*) in the simplest case. Often, a Gaussian function of Euclidean distance between points with length scale σ is used to specify connection strength:

$$w_{ij} = \begin{cases} \exp\left(-\dfrac{\left(x_i - x_j\right)^T\left(x_i - x_j\right)}{\sigma^2}\right) & \text{if } i \sim j \left('k' \text{ nearest neighbors}\right) \\ 0 & \text{otherwise} \end{cases}, \tag{1}$$

Usually, an edge $i \sim j$ is established when node i is one of k-nearest neighbors of node j or node i is within a certain Euclidean distance r, $||x_i - x_j|| < r$. The labeled nodes have labels $y_l \in \{-1, 1\} (1 = 1,\ldots, L)$, while the unlabeled nodes have zeros $y_u = 0 (u = L+1,\ldots, L+U)$. The algorithm will output an n-dimensional real-valued vector $f = [f_l^T f_u^T]^T = (f_1,\ldots, f_L f_{L+1},$

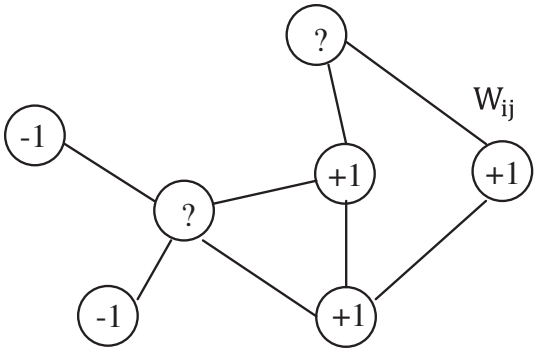

Fig. 1. Graph-based semi-supervised learning (SSL).

$\ldots f_{L+U})^T$ which can be thresholded to make label predictions on $f_{L+1},\ldots,$ f_{L+U} after learning. It is assumed that (a) f_i should be close to the given label y_i in labeled nodes and (b) overall, f_i should not be too different from its adjacent nodes f_j. One can obtain f by minimizing the following quadratic functional:

$$\mathrm{Min}_f(f-y)^T(f-y)+\mu f^T L f, \tag{2}$$

where $y=(y_1,\ldots,y_1,0,\ldots,0)^T$, and the matrix $\boldsymbol{L}$, called the graph Laplacian, is defined as $L=D-W$, $D=\mathrm{diag}(d_i)$, and $d_i=\sum_j \omega_{ij}$. The first term corresponds to the loss function in terms of condition (a), and the second term represents the smoothness of the predicted outputs in terms of condition (b). The parameter μ represents trades between loss and smoothness. The solution to Eq. (2) is obtained as

$$f=(I+\mu L)^{-1}y, \tag{3}$$

where I is the identity matrix. The formulation of Eq. (2) and its closed-form solution (Eq. (3)) present the SSL classification framework, hence the resulting thresholded value of f is ideally suited to capture the movement of oil prices.

3. Proposed method

To apply the graph-based SSL to time series prediction, we propose a method of graph representation for time series data, and a procedure for obtaining predicted values from the graph. For instance, assume that multiple time series are given as the input for the prediction problem of the WTI intermediate oil price of this month: the total amount of Saudi oil production (SAUDI), the surplus ability of OPEC production (OPEC surplus), NYMEX oil future price (NIMEX_OI), etc. To apply SSL to this problem, the proposed method begins with a re-designed graph as in Fig. 2.

The nodes in the graph represent the time series variables that influence WTI, e.g., demand- and supply-related variables and other external economic indicators (factors). Then the edge between any two nodes i~j stands for the similarity of the two sets of time series, represented as '$w_{ij}\in W$'. The label 'y_t' on each node presents either 'up' $(+1)$ or 'down' (-1) of the time series at time point t. In the graph of Fig. 2, the labels of WTI and SAUDI are not known yet at time point t, and hence are unlabeled. To estimate the label y_t, the similarity matrix of SSL was calculated at time point t-1, W_{t-1}. Based on this set-up, we explain how to measure the similarity 'w_{ij}' of a similarity matrix W and how to set the value for label 'y'.

3.1. Similarity matrix

The design of the similarity matrix W plays a critical part in the aspect of performance when using SSL [10], [44]. In the matrix W, each element represents how strongly the two nodes are related, with larger elemental value being associated with greater nodal similarity. In the proposed method, the time-series data are transformed into vectors by building technical indicators (TIs) and employing feature extraction techniques to build the similarity matrix. The general process of constructing the similarity matrix is described in Fig. 3.

3.1.1. Technical indicator (TI) transformation

TIs are frequently used in financial forecasting as they offer the advantages of removing the noise (oscillatory noise) inherent in time series and illustrating the underlying structure, i.e., the tendencies and structural factors affecting variation. Oil prices and other economic factors exist as time series data by the nature of the variables, and each of them is defined as a sequence as

$$\boldsymbol{X}_t=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_i,\ldots,\boldsymbol{x}_t\}, \tag{4}$$

where t represents the current time point, and x_t is the corresponding value. The existence of X_t as time series data induces several problems in the direct application of SSL to the data. As shown in Fig. 2, each of the nodes on the graph has its own time series, as shown in (4). For instance, the WTI node has X_t^{WTI} and the SAUDI node also has X_t^{SAUDI}.

The problem is that it is difficult to draw the similarity between them directly from the two sets of time-series data. Therefore, individual time series are transformed into structural characteristics of time point t, i.e., S_t^{WTI} and S_t^{SAUDI}, representing the trends and variations of individual series. Table 1 summarizes the TIs used in this study. The similarity between the two nodes is measured by using the seven-tuple vector $S_t=\{s_1,s_2,s_3,s_4,s_5,s_6,s_7\}$ composed of MA, BIAS, OSC, ROC, K, D, and RSI.

Using the TIs enables the time-series data to be transformed into vector-type data, while maintaining the time associations of the series, and thus eases their application to SSL. Each indicator has parameters, denoted as p or q as shown in Table 1, which must be decided by the user. However, since there is no rule for deciding appropriate parameter values, the decisions are generally made through trial-and-error. Another alternative is to consider all the diverse values of the parameters. In such a case, however, one indicator will be increased to as many variables as the number of combinations of the parameters, $p=1,\ldots,m$ and $q=1,\ldots,r$, so that the resulting vector will be represented as $S_t=\{s_1^1,\ldots,s_1^m,s_2^1,\ldots,s_2^m,s_3^{1,1},\ldots,s_3^{m,r},\ldots,s_7^1,\ldots,s_7^m\}$. This may increase the dimensions of input variables, thereby inducing the curse of dimensionality and possibly causing model over-fitting. Therefore, the next section introduces a method that uses all the diverse parameter values while reducing the unnecessary dimensions derived from the TI parameters.

3.1.2. Feature extraction

Feature extraction refers to the process of determining a mapping procedure that reduces the dimensionality and removes the noise

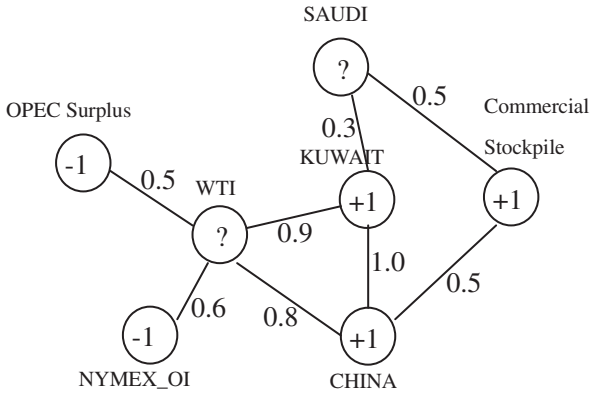

Fig. 2. Graph SSL representation for time series prediction.

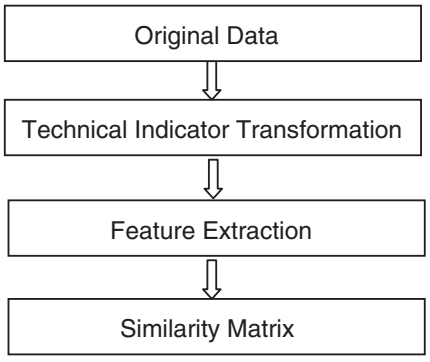

Fig. 3. Constructing the similarity matrix.

Table 1
The definition of technical indicators (TIs).

	TIs	Meaning
s_1	$MA_p(X_t) = \frac{1}{p}(x_t) + \frac{p-1}{p}MA_p(X_{t-1})$	p-moving average (exponential smoothing)
s_2	$BIAS_p(X_t) = \frac{x_t - MA_p(X_t)}{MA_p(X_t)}$	The change rate of x_t relative to $MA_p(X_t)$
s_3	$OSC_{p,q}(X_t) = \frac{MA_p(X_t) - MA_q(X_t)}{MA_p(X_t)}$	The change rate of $MA_q(X_t)$ relative to $MA_p(X_t)$
s_4	$ROC_p(X_t) = \frac{x_t - x_{t-p}}{x_t}$	The relative rate of change for X_t between p consecutive time points
s_5	$K_t^p = \frac{x_t - Min_{i=t-p-1}^{t}(x_i)}{Max_{i=t-p-1}^{t}(x_i) - Min_{i=t-p-1}^{t}(x_i)}$	Standardization of x_t
s_6	$D_t^p = MA_3(K_t^p)$	3-moving average of K_t^p
s_7	$RSI_t^p = \frac{\sum_{i=t-p-1}^{t}(\lvert x_i - x_{i-1}\rvert)}{\sum_{i=t-p-t}^{t}(\lvert x_i - x_{i-1}\rvert)}$,	The relative strength index.

effect from the data. Among the various methods for feature extraction, the linear method of principal component analysis (PCA) is the most common. By calculating the eigenvectors of the covariance matrix of original data, PCA transforms a high-dimensional input vector into a low-dimensional one whose components (extracted features) are uncorrelated. On the other hand, among the several kinds of nonlinear PCA (NLPCA), auto-associative neural network (AANN) is one of the well known nonlinear transformation methods. In AANN, the network is trained to perform identity mapping where the values of input features are approximated at the output layer, and the nonlinear principal components can be obtained from the hidden nodes in the bottleneck layer.

3.1.2.1. Principal component analysis (PCA). PCA can be used for dimensionality reduction in a data set by extracting important hidden features that provide the greatest contribution to its variance. Technically, PCA attempts to find orthonormal axes which maximally decorrelate the original features of data. Given the data points $s_i \in R^m (i = 1,\ldots,n$ and $\sum_{i=1}^{n} s_i = 1$, usually $m < n$), PCA carries out linear transformation of each s_i into a new one z_i by

$$\underbrace{Z_i}_{m\times 1} = \underbrace{U^T}_{m\times m}\underbrace{s_i}_{m\times 1}, \quad i = 1,\ldots,n, \tag{5}$$

where U is the $m \times m$ orthogonal matrix whose k^{th} column u_k is the k^{th} eigenvector of the covariance matrix $C = \frac{1}{n}\sum_{i=1}^{n} s_i s_i^T$. The matrix U can be obtained by solving the eigenvalue problem with respect to C,

$$\lambda_k u_k = C u_k, \quad k = 1,\ldots,m, \tag{6}$$

where λ_k is an eigenvalue of C and u_k is the corresponding eigenvector. The magnitude of an eigenvalue stands for the proportion of variance that can be explained by the corresponding eigenvector. Therefore, by taking the first p largest eigenvectors $\tilde{U}^T = \{u_1, u_2, \ldots, u_p\}$ we can find "lower" dimensional orthonormal space while still retaining most important aspects of the data. A projected data point onto the lower dimensional space, $\tilde{s}_i$, is calculated as the orthogonal transformations of s_i,

$$\underbrace{\tilde{s}_i}_{p\times 1} = \underbrace{\tilde{U}^T}_{p\times m}\underbrace{s_i}_{m\times 1}, \quad i = 1,\ldots,n,. \tag{7}$$

Although PCA is a well established dimensionality reduction method, its applicability is limited by the assumption that the data is a linear combination of certain features. Therefore, if the data set shows non-linear relationship among features, there is no guarantee that the extracted features by PCA will contain all important features.

3.1.2.2. Nonlinear principal component analysis (NLPCA): auto-associative neural network (AANN). Another approach to dimensionality reduction is through the use of an AANN, a special kind of feed-forward neural network [19]. AANN attempts to find and eliminate nonlinear correlations in the data. Similar to PCA, it can be used to reduce the dimensionality of data by removing redundant features. The general structure of AANN, as shown in Fig. 4, consists of an input layer, an output layer, and multiple hidden layers. Both the number of input nodes and that of output nodes are equally set to m. Among the hidden layers, the mapping layer models the mapping function (F1) and the demapping layer models the demapping function (F2). The number of nodes, p, in a particular hidden layer, the so called "bottleneck layer", is set to be less than the number of nodes in the input/output layer ($p < m$). In auto-associative mapping, the target data is set to be identical to the input data. This "identity mapping" creates a global reduction of the data dimensionality while the input data goes through the bottleneck layer before appearing at the output layer. Let F denote the auto-associative mapping learnt by the network. If $\{\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_n\}$ is the set of output data produced by the AANN when the input data set $\{s_1, s_2, \ldots, s_n\}$ is given, then F can be found which minimizes the mean square error,

$$E = \sum_{i=1}^{n}\left(s_i - \tilde{s}_i\right)^T\left(s_i - \tilde{s}_i\right) = \sum_{i=1}^{n}(s_i - F(s_i))^T(x_i - F(s_i)). \tag{8}$$

The mapping function F can be separated into F_1 and F_2, so that $F(.) = F_2(F_1(.))$, where F_1 is the transformation in the network from the input layer into the dimension compressing the hidden layer (the bottleneck layer), and F_2 is the transformation from the bottleneck layer into the output layer. To summarize, the data are first compressed to lower the dimensionality and then reconstructed. The mapping from the input layer to the bottleneck layer can be regarded as a "nonlinear" projection onto the lower dimensional space ($m \to p$), and each node in the bottleneck can be considered as an extracted feature retaining significant information of the data. New data $\tilde{s}_i$ is then calculated as

$$\underbrace{\tilde{s}_i}_{p\times 1} = F_1\underbrace{(s_i)}_{m\times 1}. \tag{9}$$

AANN is a good model to extract the variables that can well express the nonlinear relationship of data if its structure is well established. However, since AANN is not a method in which the number of the nodes of the bottleneck layer is determined from the beginning, its need to be determined by the users in accordance with the situation introduces significant difficulty.

3.2. Label

The label on the node in the SSL graph in Fig. 2 is designed to explain whether the predicted value of the corresponding variable is up or down. It can be formulated as follows:

$$y_t = \text{sign}(x_t - MA_3(x_t)). \tag{10}$$

For instance, if the total amount of SAUDI oil production of this month (t) exceeds its three-month moving average, Eq. (10) will give a '$y_t = +1$' label. On the contrary, the node is labeled as '$y_t = -1$' for the opposite case. And '$y_t = 0$' if there is no information about the movement of the corresponding time-series value at time point t; the label is to be predicted. In the proposed method, we set the label of the target variable, WTI spot prices, to '0'. Also note that some of input variables may not be able to be labeled, for instance, SAUDI in Fig. 2, but SSL produces a prediction even for such a node.

Given label y_t, Eq. (3) provides the predicted value f_t for every node, which can take on a real number unlike the values of label y_t. The following interpretation can be put on the predicted value. At time point t, if the predicted value f_t is an arbitrary positive number, then it is equivalent in sign with the value of $(x_t - MA_3(x_t))$. This

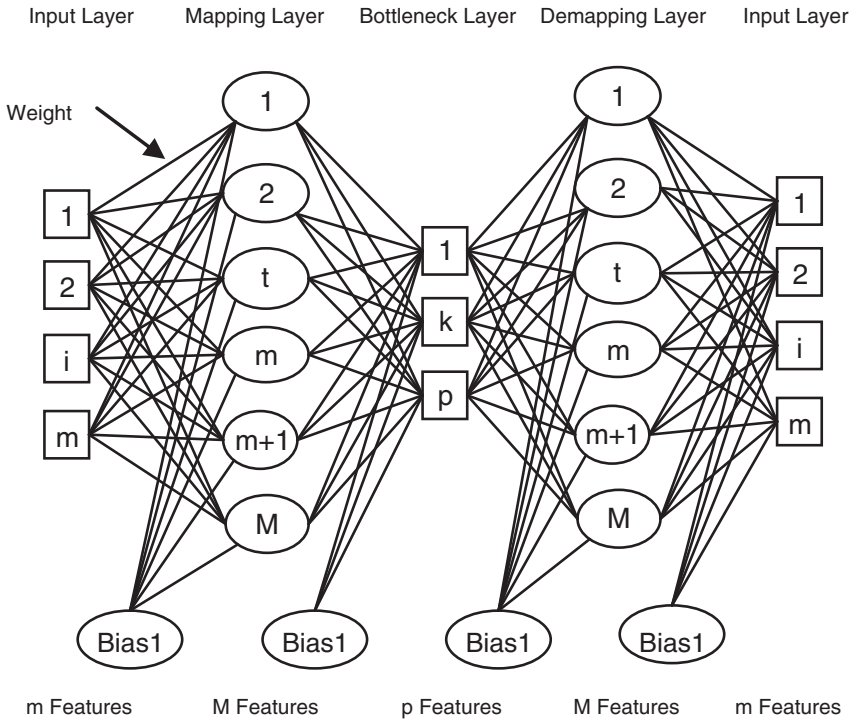

Fig. 4. Architecture of AANN.

means that x_t, i.e., the WTI oil price for time point *t* will exceed the price of the past three-month moving average. To rephrase this

$$\text{sign}(f_t) > 0 \Leftrightarrow \text{sign}(x_t - MA_3(x_t)) > 0 \tag{11}$$

and thus, the following inequality holds:

$$x_t > MA_3(x_t). \tag{12}$$

To substitute the right-hand side of the inequality with the definition of the moving average equation in Table 1, the following inequality is derived

$$x_t > \frac{1}{3}x_t + \frac{3-1}{3}MA_3(x_{t-1}),$$

and summarized as the final form below.

$$x_t > MA_3(x_{t-1}) \tag{13}$$

That is, since '$f_t > 0$' actually means $x_t > MA_3(x_{t-1})$, this implies that 'one-time-point-ahead prediction' is available with the predicted value f_t and $MA_3(x_{t-1})$ at time point t-1. This procedure is schematized as shown in Fig. 5.

4. Experiment

In this section, we implement the proposed method on the prediction of price movement of West Texas Intermediate crude oil. As aforementioned in the earlier sections, a set of multiple time series data is described using a network (or a graph) to capture the multiple interactions included and prediction is made using SSL. Since the construction procedure of the network employs TI transformation which increases dimensionality, feature extraction is implemented through alternative approaches of PCA and NLPCA. Depending on the number of extracted features, many models are possible, and therefore the best model is determined by performance comparison using a measure known as the area under the ROC curve (AUC). Finally, comparison with other representative data mining models is made of the best SSL model determined previously.

4.1. Data

The data employed in this study are the time-series data of the prices of the WTI crude oil which consist of 198 monthly spot prices ranging from January 1992 to June 2008 as shown in Fig. 6.

For the same span, the Korea Energy Economics Institute (KEEI) made available the 25 diverse external economic factors (time series variables) that are strongly related to the WTI oil price screened by the domain experts, e.g., demand- and supply-related variables and other external economic indicators. The demand-related variables include the amount of overall international oil production, the amount of OPEC oil production, and the amount of SAUDI oil production. The supply-related variables include the amount of overall international demand, the amount of OECD countries' oil demand, and the amount of non-OECD countries' oil demand. Other economic indicators include the producer price indices and the US dollar exchange rates. We tested the association of those 25 input variables with the WTI oil price in the continuous scale with R^2 improvement values at the

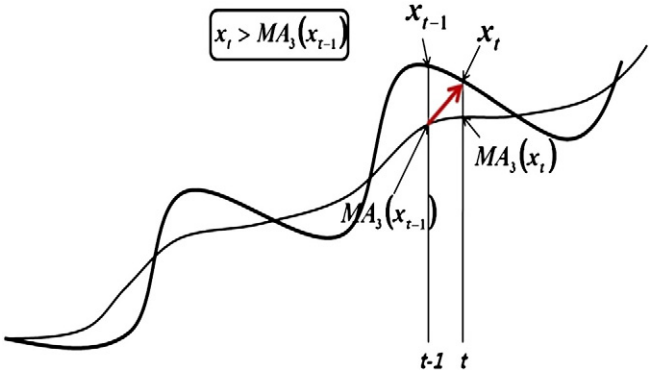

Fig. 5. Schematic description of interpreting forecasted values (when $f_t > 0$).

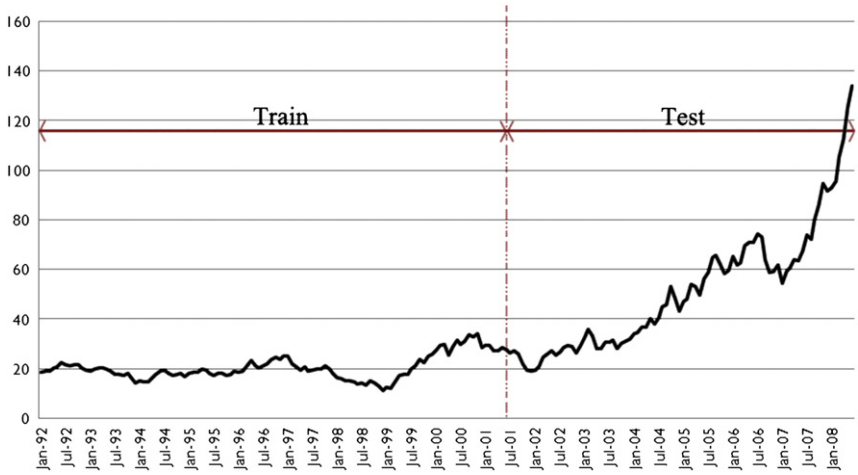

Fig. 6. The monthly WTI crude oil prices from Jan. 1992 to Jun. 2008.

significance level of $\alpha = 0.01$ in order to reject the insignificant ones [29]. But all of the given 25 variables showed statistical significance, and hence employed all as input variables in our study. The variables are tabulated in Table 2. Among them, WTI is the target variable and so regarded as unlabeled. Fig. 7 shows how the 25 external economic factors used as input variables affect WTI. Notably, input variables not only affect WTI but also affect each other in a very complicated manner.

The data set used in the experiment was set up as follows. The first 100 monthly data values for January 1993 through April 2001 were used as a training data set. The remaining 86 monthly data values for May 2001 through June 2008 were used as a test set (of the 198 time points in total, 12 time points necessary to create TIs were excluded) for performance comparison with competing models.

Table 2
The 26 sets of time-series data from January 1992 to June 2008.

Target variable		West Texas intermediate crude oil prices (WTI)	
Input variables		Association with target variable	
		R^2	*p*-value
Demand-side	Overall amount of world oil demand	0.941	1.12×10^{-41}
	Amount of OECD demand	0.373	1.22×10^{-12}
	Non-OECD demand	0.605	3.86×10^{-65}
	China demand	0.227	6.83×10^{-59}
	USA demand	0.365	3.32×10^{-21}
Supply-side	OPEC production	0.773	3.49×10^{-59}
	Saudi production	0.736	1.67×10^{-39}
	Iran production	0.630	2.33×10^{-40}
	Iraq production	0.738	1.02×10^{-8}
	Kuwait production	0.585	2.32×10^{-30}
	Non-OPEC production	0.593	6.57×10^{-29}
	USA production	0.554	5.57×10^{-5}
	Russia production	0.487	2.67×10^{-48}
	World production	0.469	2.23×10^{-44}
Other economic indicators	Producer price index	0.259	2.98×10^{-12}
	U.S. exchange rate	0.663	8.64×10^{-22}
	OECD commercial stockpiles	0.669	4.83×10^{-15}
	U.S. commercial stockpiles for crude oil	0.326	5.44×10^{-5}
	U.S. commercial stockpiles for oil	0.319	2.29×10^{-5}
	OPEC surplus production ability	0.206	1.66×10^{-11}
	NYMEX oil futures price	0.824	2.37×10^{-76}
	Non-commercial real purchase (short)	0.133	1.03×10^{-4}
	Non-commercial real purchase (long)	0.194	7.48×10^{-11}
	Commercial volume (short)	0.048	1.91×10^{-4}
	Commercial volume (long)	0.031	1.30×10^{-4}

4.2. Performance measure (AUC)

To measure the prediction performance, the area under the curve (AUC), which is defined as the area under the receiver operating characteristic (ROC) curve, is used [14,15]. The ROC curve plots true positive rate as a function of false positive rate for differing classification thresholds as shown in Fig. 8. The AUC measures the overall quality of the ranking induced by model rather than the quality of a single value of threshold in that ranking. The closer the curve follows the left-hand border and then the top-border of the ROC space, the larger value of AUC the model produces; i.e., the more accurate the model is.

4.3. SSL parameter selection

The parameter values of the SSL model, k and μ, the number of k-nearest neighbors in (1) and the loss-smoothness tradeoff in (3) were selected from $\{k,\mu\}\in\{2,3,4,5\}\times\{0.01,0.1,0.3,0.5,0.7,1,10,100\}$ as optimum combinations through cross-validations. Fig. 9 illustrates a typical pattern of how the AUC performance of an SSL model would vary depending on the combinations of parameters k and μ. Every SSL model in this study found its model-parameters at its best AUC, for instance, the model in Fig. 9 set the values of (k, μ) to (2, 0.1).

4.4. Results on TI transformation and feature extraction

As shown in Section 3.1.1, each of the 26 variables is transformed into the seven TIs: MA, BIAS, OSC, ROC, K, D or RSI. The parameters for each TI are set as $p\in\{3,4,6,8,9,12\}$. Since a single variable is transformed to 7 TIs, and each of which has 6 dependent sub-variables, then the total number of the sub-variables per variable, or simply input dimensionality, becomes 42 (= 7 TIs × 6 parameter dependent sub-variables). Even though the use of TI facilitates the consideration of the trends and the structure of the data, there is, on the other hand, the drawback that one variable turns into a set of an increased number of sub-variables. The increased number of input variables means an increase in dimensionality, which degrades the performance of the prediction model. Thus, as mentioned in Section 3.1.2, feature extraction techniques are employed: PCA and NLPCA. If we extract a single feature per TI, then the 6 parameter dependent sub-variables are reduced to one dimensional feature. For PCA, this implies that we use only the first principal component from the covariance matrix of the 6 parameter dependent sub-variables. For NLPCA, the network configuration is composed of 6 input nodes in the input layer, 1 hidden node in the bottleneck layer, and 6 output nodes in the output layer (See the architecture of AANN in Fig. 4). During the network training, the values of the 6 parameter dependent sub-variables are fed to the network as both input and target.

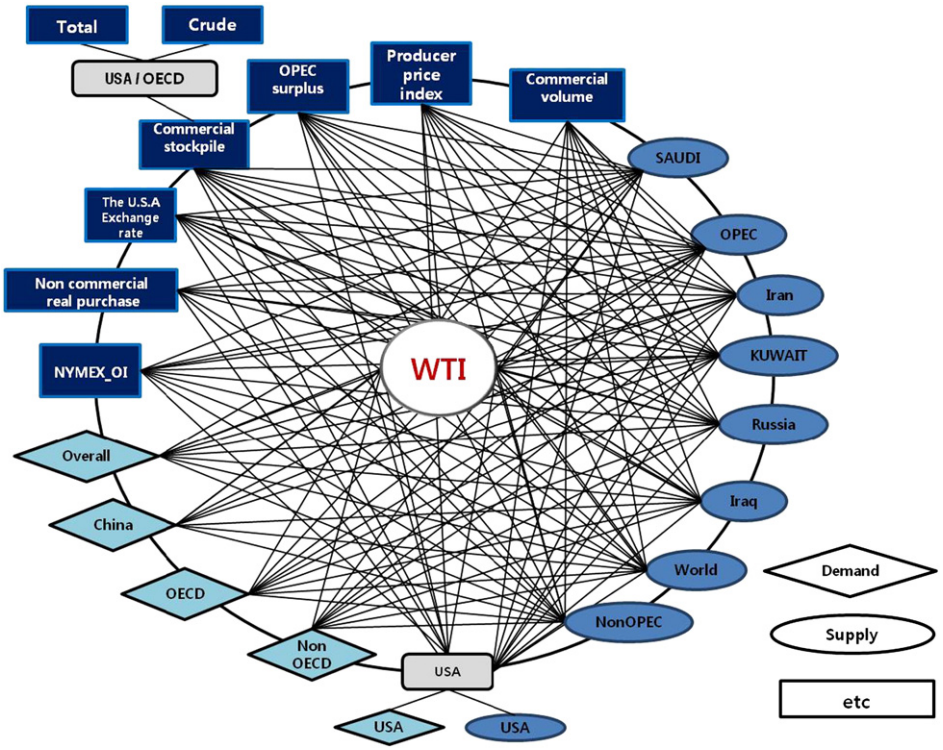

Fig. 7. The SSL graph for the 26 sets of time-series variables: WTI is the target variable and the remaining economic factors are input variables. The edge between two nodes represents the similarity between them.

After training, the output value of the bottleneck node is used as the extracted feature. Fig. 10 shows the procedure of feature extraction following TI transformation from a variable, SAUDI.

This process is similarly applied to all of the 26 variables. Then a node in the graph which corresponds to a variable is represented as a vector of 7 tuples, e.g., $S^{SAUDI} = (s_1^{SAUDI}, s_2^{SAUDI}, s_3^{SAUDI}, s_4^{SAUDI}, s_5^{SAUDI}, s_6^{SAUDI}, s_7^{SAUDI})$. The similarity (edge-connection) between the nodes is built from (1). Fig. 11 exemplifies how the connections between the 26 variables are made from 7-tuple vector-representation. The size of vector can vary depending on how many features are extracted. If we set the number of extracted features to three, then a variable is represented as a vector of 21 tuples, $S = (s_{11}, s_{12}, s_{13}, s_{21}, s_{22}, s_{23}, \ldots, s_{51}, s_{52}, s_{53}, \ldots, s_{71}, s_{72}, s_{73})$. However, it is difficult to determine the number of extracted features when the intrinsic dimension is unknown. In the experiment, we attempted to find the optimum number of extracted features among 1, 3 and 6, which respectively led to 7, 21 and 42 input dimensionality per variable (=7 TIs × {1, 3, 6} features extracted from 6 parameter dependent sub-variables).

In the following Fig. 12 and Table 3, the AUC values of the seven SSL models are compared: SSL_0, SSL_{P1}, SSL_{P3}, SSL_{P6}, SSL_{N1}, SSL_{N3} and SSL_{N6}. The designations were determined based on whether PCA or NLPCA was used and the number of extracted features. For instance, SSL_{P3} means a model made by extracting three features per TI through PCA and then applying SSL. The input dimensionality is 21 (=7 TIs × 1 extracted feature). Likewise, SSL_{N3} is a model made by extracting three features through NLPCA. The model designated as SSL_0 refers to a model without the feature extraction procedure, therefore 6 sub-variables per TI are all used. The input dimensionality of SSL_0 is 42 (=7 TIs × 6 parameter dependent sub-variables).

The average AUC of SSL_0 using 42 sub-variables is 0.84. A notable fact is that SSL_{P1} almost reproduces the performance of SSL_0 with only 7 features with the average AUC of 0.83. Overall, the performances of the

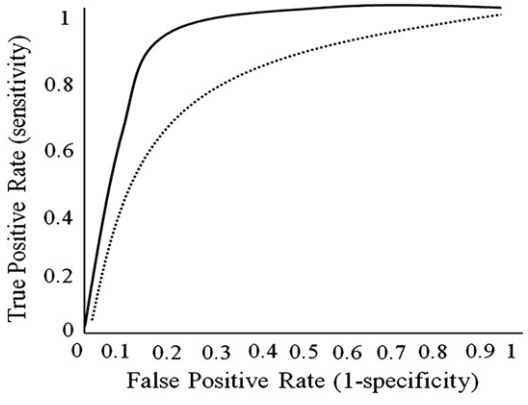

Fig. 8. ROC curve.

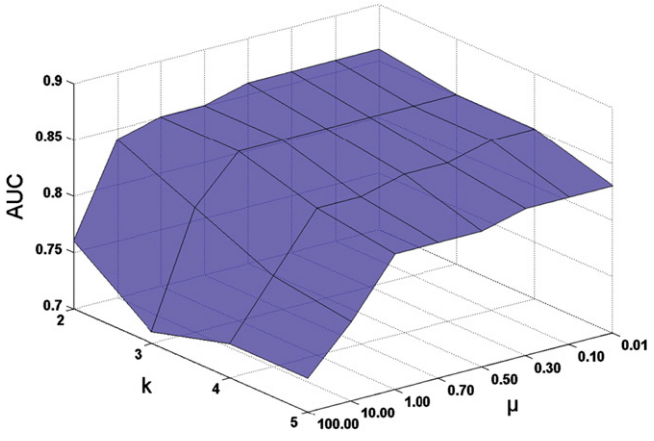

Fig. 9. The AUC over model-parameters variation (k and μ) using the SSL model.

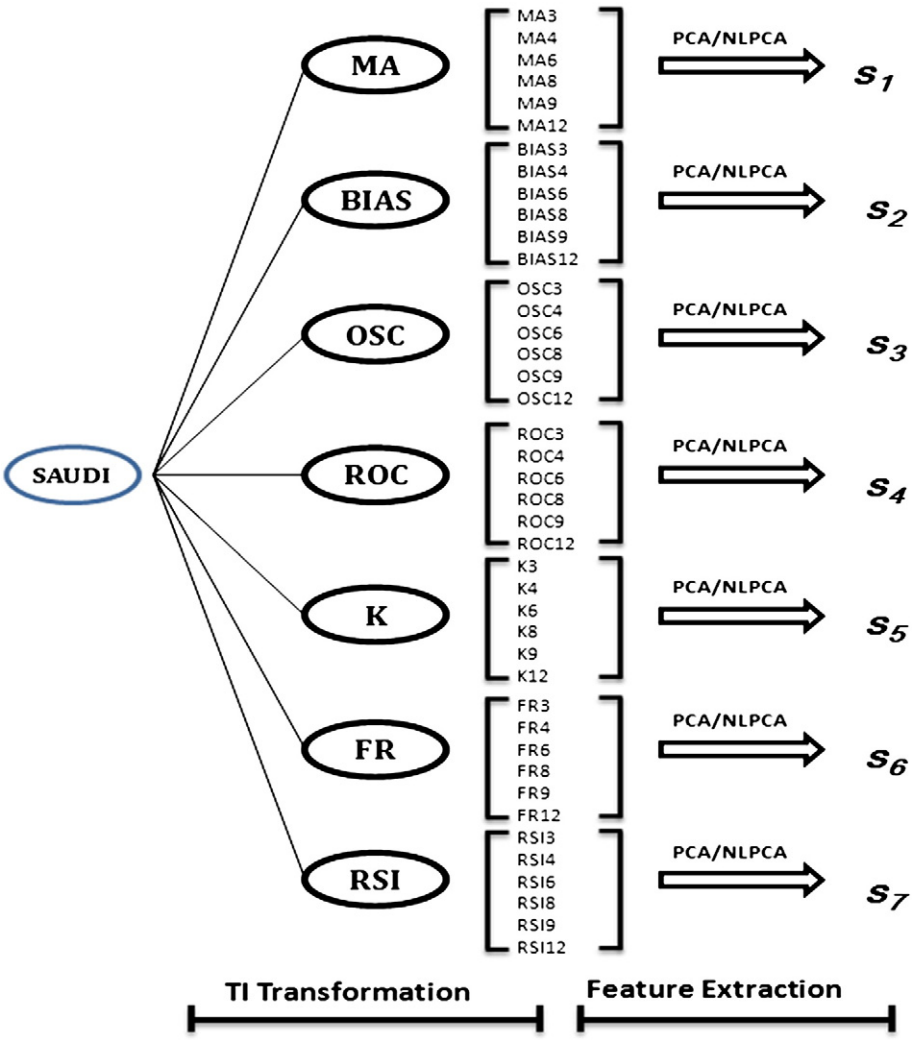

Fig. 10. The procedure of feature extraction following TI transformation.

SSL_P models are similar to the performance of the SSL_0 models. The performances of the SSL_N tend to be somewhat inferior. The results that the SSL_P models are better performed than the SSL_N models give us a hint that the 6 parameter dependent variables would be rather linearly correlated since they are derived from an identical TI formula, e.g. MA(p) where p = {3,4,6,8,9,12}, and also see Fig. 10. Among the SSL_P models, SSL_{P3} shows the best performance with an average AUC of 0.86. The optimum number of extracted features is usually determined by a trial-and-error fashion. However, related to our experimental setting for the number of extracted features {1, 3, 6}, we may conjecture that a single feature would not be sufficient to explain the variability among the 6 parameter dependent variables whereas 6 features would just reconstruct the original input space into the feature space without the effect of dimensionality reduction and thus no effect of noise reduction. Using 3 extracted features would be a compromise between both ends. To summarize, if feature extraction is used, performances similar or superior to the original performance can be expected even with smaller numbers of variables and, in particular, PCA was more effective than NLPCA in our experiment.

4.5. Results of the comparison: SSL vs. other models

In this experiment, SSL_{P3} was compared with five well known representative models: an auto-regression (AR) model, a logistic regression (LR) model, an ANN model, and two SVM models, $SVM_{\_RBF}$ and $SVM_{\_POLY}$ using RBF kernel function and polynomial kernel function, respectively. The optimum model parameters were selected for each of the five models in Section 4.2 in a similar way to that done for SSL. The resultant AUC values of the six models are summarized in Table 4 and Fig. 13. First, AR and LR showed average AUC values of 0.53 and 0.55, respectively, which are much smaller than the AUC average of 0.86 for the proposed SSL_{P3}. This indicates that the time-series models, based on existing linear models, have limitations in explaining the irregular patterns of oil price movement. Although ANN and SVM showed average AUC values of 0.74 and 0.66, respectively, which indicated their superior accuracy compared to that of AR and the LR, they showed relatively poor results compared to SSL_{P3}. The superior generalization ability of SSL compared to that of ANN and SVM was attributed to the fact that SSL uses not only one-to-one relationships between the target variable WTI and the input variables (demands, supplies and other external economic factors) but also the intrinsic inter-relationships between the input variables. This enabled the SSL to perform more accurately than others. Fig. 14 presents how SSL_{P3} fits the ups and downs of the WTI oil prices during the test period of May 2001 through June 2008. The thicker line indicates the WTI oil prices and the thinner line its three-month moving average, i.e., MA_3(WTI). The figure shows that the predicted value fits the ups and downs of the price movement reasonably well.

5. Conclusions

This paper has proposed a novel method for oil price prediction using the SSL algorithm. The proposed method modifies the existing

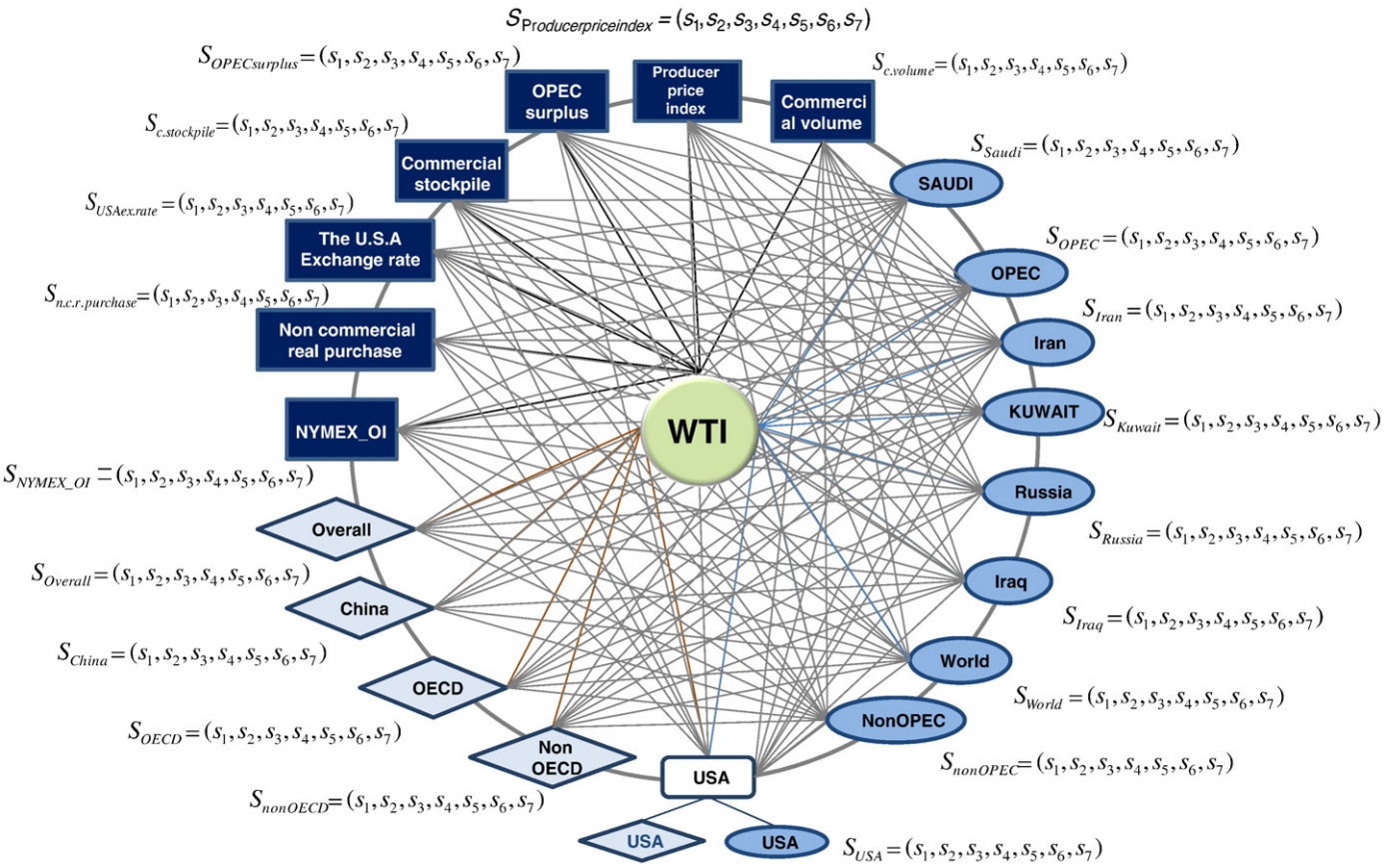

Fig. 11. Similarity (edge-connection) calculation from 7-tuple vector-representation.

SSL algorithm for application to time-series prediction, including measuring the similarity between different sets of time-series data and the labels of pricing ups and downs, and the advanced techniques using TI transformation and feature extraction. The advantages of the proposed method can be summarized as follows. First, the modified SSL considers not only the influence of input variables on the target variable but also the mutual influences among the input variables. For our oil price prediction problem, the WTI crude oil prices were predicted by taking into account the influence of external economic factors such as demand-side factors, supply-side factors, and various types of international economic index. The economic factors including the oil price were represented as nodes in a network, and connected via similarities between them. Then prediction on the WTI crude oil price was made by the propagated influence of its neighboring economic factors through the connections. This enables to resolve the complexity and irregularity of oil price prediction problem caused by its intrinsic dynamics interacting with many global or national economic factors, and results in more accurate prediction than that offered by the existing representative prediction models. Second, by transforming time-series data into TIs, the noise in the data was removed and the underlying tendencies and structural factors for the variations were revealed. Third, by using feature extraction, only the few features that are commonly intrinsic among input variables were used in the modeling, which avoided any unnecessary increases of the input dimensionality. The synergy effect of these three advantages were harmonized in our SSL model-based oil price prediction, and afforded an AUC accuracy of 0.86, which is an unprecedented performance. The proposed method is expected to be applied to any domain that requires time-series prediction, i.e., the prediction of international oil prices, domestic/foreign stock price indices, price variability, national growth rates and currency exchange rates. Technically, the proposed method can be more sophisticated with respect

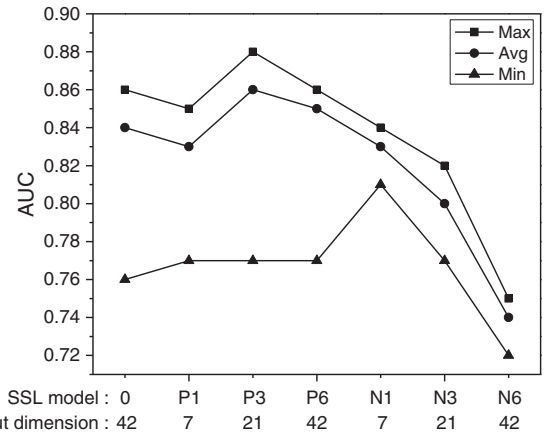

Fig. 12. Comparison of AUC for seven different SSL models. The numbers under the graph stand for the number of the extracted features through PCA or NLPCA. The squares indicate the best AUC after repetition of experiments under every combination of parameters $\{k,\mu\}$, the circles indicate the average AUC and the triangles indicate the minimum AUC.

Table 3
AUCs summary for seven different SSL models.

AUC	Max	Avg	Min
SSL_0	0.86	0.84	0.76
SSL_{P1}	0.85	0.83	0.77
SSL_{P3}	0.88	0.86	0.77
SSL_{P6}	0.86	0.85	0.77
SSL_{N1}	0.84	0.83	0.81
SSL_{N3}	0.82	0.80	0.77
SSL_{N6}	0.75	0.74	0.72

Table 4
AUCs comparison of SSL vs. the five competing models.

AUC	Max	Avg	Min	Rank
SSL_{P3}	0.88	0.86	0.77	1
AR	0.54	0.53	0.52	6
LR	0.64	0.55	0.49	5
ANN	0.82	0.74	0.55	2
$SVM_{\_RBF}$	0.78	0.73	0.67	3
$SVM_{\_POLY}$	0.74	0.58	0.50	4

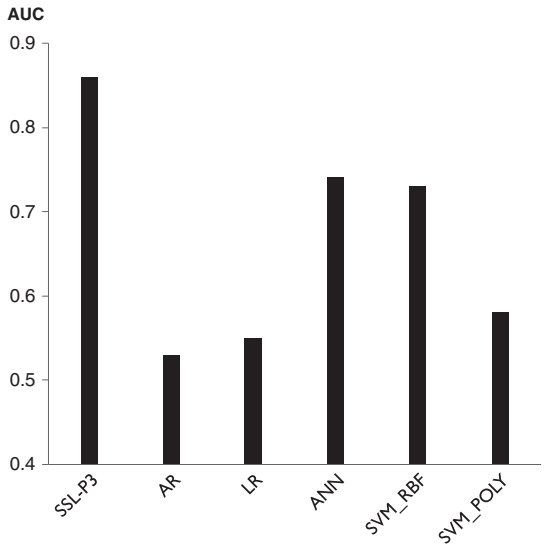

Fig. 13. Comparison in AUC: SSL versus the five competing models.

to feature extraction procedure. PCA/NLPCA can be replaced with independent component analysis (ICA) as in [23], and a combining approach of multiple features can be an alternative of choosing one of them [36]. Applying and adapting our method to diverse domains and techniques will be well worth further research.

Acknowledgment

The authors would like to gratefully acknowledge support from Post Brain Korea 21 and the research grant from National Research Foundation of the Korean Government (2010-0007804/2012-0000994).

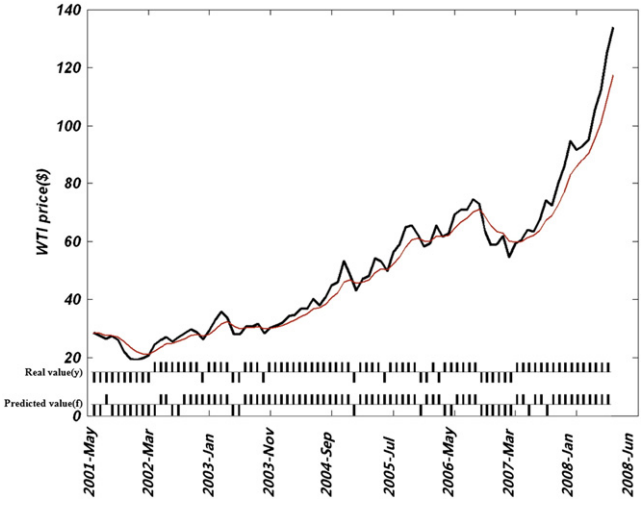

Fig. 14. SSL_{P3} for the test period of May 2001 through Jun 2008.

References

[1] S. Abosedra, H. Baghestani, On the predictive accuracy of crude oil futures prices, Energy Policy 32 (2004) 1389–1393.
[2] A.T. Akarca, D. Andrianacos, Detecting break in oil price series using the Box–Tiao method, International Advances in Economic Research 3 (1997) 217–224.
[3] R.A. Amano, S.V. Norden, Exchange rates and oil prices, Review of International Economics 6 (1998) 683–694.
[4] R.K. Ando, T. Zhang, A High-Performance Semi-Supervised Learning Method for Text Chunking, in: ACL '05 Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics Ann Arbor, Michigan, 2005, pp. 1–9.
[5] E. Bair, R. Tibshirani, Semi-supervised methods to predict patient survival from gene expression data, PLoS Biology 2 (2004) 511–522.
[6] S.A. Basher, P. Sadorsky, Oil price risk and emerging stock markets, Global Finance Journal 17 (2004) 224–251.
[7] F. Birol, Analysis of the Impact of High Oil Price on the Global Economy, International Energy Agency, 2004.
[8] A. Blum, S. Chawla, Learning from labeled and unlabeled data using graph mincuts, in: ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning San Francisco, 2001, pp. 19–26.
[9] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: COLT' 98 Proceedings of the eleventh annual conference on Computational learning theory New York, 1998, pp. 92–100.
[10] O. Chapelle, B. Scholkopf, A. Zien, Semi-Supervised Learning Cambridge, MIT Press, England, 2006.
[11] G. Cortazar, E.S. Schwartz, Implementing a stochastic model for oil futures prices, Energy Economics 25 (2003) 215–238.
[12] L. Feng, J. Li, X. Pang, China's oil reserve forecast and analysis based on peak oil models, Energy Policy 36 (2008) 4149–4153.
[13] Y.-C. Gong, C.-L. Chen, Semi-supervised method for gene expression data classification with Gaussian fields and harmonic functions, in: 19th International Conference on Pattern Recognition (ICPR 2008), Tampa, FL, 2008, pp. 1–4.
[14] M. Gribskov, N.L. Robinson, The use of receiver operating characteristic (ROC) analysis to evaluate sequence matching, Computers and Chemistry 20 (1996) 25–33.
[15] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic, Radiology 143 (1982) 29–36.
[16] L.-Y. He, Y. Fan, Y.-M. Wei, Impact of speculator's expectations of returns and time scales of investment on crude oil price behaviors, Energy Economics 31 (2009) 77–84.
[17] D. Huang, B. Yu, F.J. Fabozzi, M. Fukushima, CAViaR-based forecast for oil price risk, Energy Economics 31 (2009) 511–518.
[18] T. Joachims, Transductive inference for text classification using support vector machines, in: International Conference on Machine Learning, San Francisco, 1999, pp. 200–209.
[19] R.T. Kamimura, S. Bicciato, H. Shimizu, J. Alford, G.N. Stephanopoulos, Mining of multivariate temporal biological data: a framework for the rational design of data-driven models, in: Presented at the BioKDD, 2001: Workshop on Data Mining in Bioinformatics, San Francisco, CA(US), 2001.
[20] T.A. Knetsch, Forecasting the price of crude oil via convenience yield predictions, Journal of Forecasting 26 (2007) 527–549.
[21] A. Lanza, M. Manera, M. Giovannini, Modeling and forecasting cointegrated relationships among heavy oil and product prices, Energy Economics 27 (2005) 831–848.
[22] R. Liu, J. Zhou, M. Liu, A graph-based semi-supervised learning algorithm for web page classification, in: International Conference on Intelligent Systems Design and Applications, China, 2006, pp. 856–860.
[23] C.-J. Lu, T.-S. Lee, C.-C. Chiu, Financial time series forecasting using independent component analysis and support vector regression, Decision Support Systems 47 (2009) 115–125.
[24] S. Mirmirani, H.C. Li, A comparison of VAR and neural networks with genetic algorithm in forecasting price of oil, Applications of Artificial Intelligence in Finance and Economics 19 (2004) 203–223.
[25] C. Morana, A semiparametric approach to short-term oil price forecasting, Energy Economics 23 (2001) 325–338.
[26] K. Nigam, A.K. Mccallum, S. Thrun, T. Mitchell, Text classication from labeled and unlabeled documents using EM, Machine Learning 39 (1999) 1–34.
[27] J. Philip, K. Verleger, Adjusting to Volatile Energy Prices, Policy Analyses in International Economics Series, vol. 39, Peterson Institute, 1994. (ISBN: 0881320692, 9780881320695)
[28] SajalGhosh, Import demand of crude oil and economic growth: evidence from India, Energy Policy 37 (2009) 699–702.
[29] K.S. Sarma, Variable selection node, in: Predictive modeling with SAS Enterprise Miner: practical solutions for business, ed Cary, NC, USA: SAS Institute Inc, 2007, pp. 48-50.
[30] H. Shin, K. Tsuda, Prediction of protein function from networks, in: O. Chapelle, et al., (Eds.), Semi-Supervised Learning, MIT Press, 2006, pp. 339–352.
[31] H. Shin, A.M. Lisewski, O. Lichtarge, Graph sharpening plus graph integration: a synergy that improves protein functional classification, Bioinformatics 23 (2007) 3217–3224.
[32] H. Shin, K. Tsuda, B. Schoelkopf, Protein functional class prediction with a combined graph, Expert Systems with Applications 36 (2) (November 2009) 3284–3292.
[33] H. Shin, N.J. Hill, A.M. Lisewski, J.-S. Park, Graph sharpening, Expert Systems with Applications 37 (2010) 7870–7879.
[34] P. Stevens, The determination of oil prices 1945–1995: a diagrammatic interpretation, Energy Policy 23 (1995) 861–870.
[35] A. Subramanya, J. Bilmes, Soft-supervised learning for text classification, in: EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing Honolulu, Hawaii, 2008, pp. 1090–1099.

[36] C.-F. Tsai, Y.-C. Hsiao, Combining multiple feature selection methods for stock prediction: union, intersection, and multi-intersection approaches, Decision Support Systems 50 (2010) 258–269.
[37] W. Xie, L. Yu, S. Xu, S. Wang, A new method for crude oil price forecasting based on support vector machines, in: Presented at the International Conference on Computational Science, 2006.
[38] D. Yarowsky, Unsupervised word sense disambiguation rivaling supervised methods, in: ACL '95 Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics Stroudsburg, 1995, pp. 189–196.
[39] S. Yousefi, I. Weinreich, D. Reinarz, Wavelet-based prediction of oil prices, Chaos, Solitions and Fractals 25 (2005) 265–275.
[40] L. Yu, S. Wang, K.K. Lai, Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm, Energy Economics 30 (2008) 2623–2635.
[41] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, B. Schölkopf, Learning with local and global consistency, in: Advances in Neural Information Processing Systems 16(NIPS), Whistler, Britishi Columbia, 2004, pp. 321–328.
[42] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, B. Schölkopf, Learning with local and global consistency, Advances in Neural Information Processing Systems 16 (2004) 321–328.
[43] X. Zhu, Semi-supervised learning with graphs, PhD thesis, Carnegie Mellon University, CMU-LTI-05–192, 2005.
[44] X. Zhu, Semi-supervised learning literature survey, Technical Report 1530, Computer Science, University of Wisconsin-Madison, 2005.
[45] X. Zhu, Z. Ghahramani, J. Lafferty, Semi-supervised learning using Gaussian fields and harmonic functions, in: International Conference on Machine Learning (ICML2003), Washington DC, 2003, pp. 912–919.

Hyunjung (Helen) Shin received the Ph.D. degree in Data Mining from Seoul National University, and further majored in Machine Learning during her Post-Doc at Max Planck Institute in Germany. Since 2006, she joined Ajou University as a faculty member of the Department of Industrial and Information Systems Engineering.
Theory interest of her is more focused on Data Mining algorithms including Machine Learning.
Her research activities on application range across areas as different as hospital fraud detection, direct marketing in CRM, Oil/Stock price prediction, bio-medical informatics, etc.

Tianya Hou received her B.E. degree from Tsinghua university in China and further majored Data Mining as her M.S. degree in Ajou University in South Korea. Her current research topic is oil price prediction using neural network and semi-supervised learning.

Kanghee Park received B.E. degree from Ajou University in 2008, and is currently pursuing his Ph.D. degree at Graduate School of Industrial Engineering, Ajou University, South Korea. His research interest is on financial time-series prediction using various techniques of machine learning algorithms.

Chan-Kyoo Park is an Associate Professor of School of Business at Dongguk University in Seoul. He earned his Ph.D. in Operations Research from Department of Industrial Engineering at Seoul National University. His current research interests are in the areas of management science and data mining. His research has been published in several journals including European Journal of Operational Research, Computers and Operations Research, and Asia-Pacific Operations Research.

Sunghee Choi earned his Ph.D. in Economics at the Claremont Graduate University of Claremont Colleges, USA. He is an assistant professor of International Commerce at the Keimyung University, Rep. of Korea.