Energy Economics 66 (2017) 9–16

Contents lists available at ScienceDirect

Energy Economics

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

A deep learning ensemble approach for crude oil price forecasting

Yang Zhao [a,b], Jianping Li [a,b,*], Lean Yu [c]

[a] Institute of Policy and Management, Chinese Academy of Sciences, Beijing, China
[b] University of Chinese Academy of Sciences, Beijing, China
[c] School of Economics and Management, Beijing University of Chemical Technology, Beijing, China

ARTICLE INFO

Article history:
Received 7 December 2016
Received in revised form 27 March 2017
Accepted 29 May 2017
Available online 09 June 2017

JEL classification:
C45
C52
C53
Q47
E37
C63

Keywords:
Crude oil price forecasting
Deep learning
Ensemble learning
Stacked denoising autoencoder
Bagging
Multivariate forecasting

ABSTRACT

As crude oil price is influenced by numerous factors, capturing its behavior precisely is quite challenging, and thus leads to the difficulty of forecasting. In this study, a deep learning ensemble approach is proposed to deal with this problem. In our approach, two techniques are utilized. One is an advanced deep neural network model named stacked denoising autoencoders (SDAE) which is used to model the nonlinear and complex relationships of oil price with its factors. The other is a powerful ensemble method named bootstrap aggregation (bagging) which generates multiple data sets for training a set of base models (SDAEs). Our approach combines the merits of these two techniques and is especially suitable for oil price forecasting. In the empirical study, the WTI crude oil price series are investigated and 198 economic series are used as exogenous variables. Our approach is tested against some competing approaches and shows superior forecasting ability that is statistically proved by three tests.

1. Introduction

Crude oil price volatilities have great impact on the economic activities of the world from many aspects. It has been largely free to fluctuate in response to the forces of supply and demand (Baumeister and Kilian, 2016). Besides these two fundamentals, various factors strike the oil prices at different frequencies. In energy market, production of other commodities including natural gas, coal and renewable energy, may have substitution effect which leads to the volatility of oil price indirectly. Other factors such as financial markets, economic growth, technology development and irregular events also influence the oil price in different ways. Complex relationships are built between these factors and oil prices, thus drive strong fluctuations in crude oil market. As a result, forecasting oil price has always been a tough task. However, seeking for promising forecasting approaches for oil price series is hardly outdated since crude oil is the main source of energy in the world and dominates the economic activities. Accurate forecast of oil price guides the decision making of many sectors such as business organizations and governments.

Research on crude oil price forecasting has lasted for decades and plenteous approaches have been proposed. Besides the classic econometric approaches, various machine learning methods are utilized to mining the inner complexity of oil price. The most typical and commonly used machine learning methods include neural network (NN) and support vector machine (SVM). They are particularly welcomed for their capability of modeling complex characteristics such as nonlinearity and volatility. For example, genetic algorithm (GA) (Kaboudan, 2001), NN (Moshiri and Foroutan, 2006) and SVM (Xie et al., 2006) are first applied to forecast crude oil price in earlier studies. In recent years, semi-supervised learning (SSL) (Shin et al., 2013), gene expression programming (GEP) (Mostafa and El-Masry, 2016) are used for oil price forecasting. The above mentioned models are single models of original form while hybridization of single machine learning models (especially NN) for oil price forecasting is becoming a popular phenomenon. Hybrid models have proved to have better forecasting accuracies than their corresponding single machine learning techniques. For example, Godarzi et al. (2014) forecast oil price with an NN based dynamic

* Corresponding author at: Institute of Policy and Management, Chinese Academy of Sciences, Beijing 100190, China.
E-mail address: ljp@casipm.ac.cn (J. Li).

http://dx.doi.org/10.1016/j.eneco.2017.05.023
0140-9883/

nonlinear autoregressive with exogenous input (NARX) model which is a multivariate forecasting model. Yu et al. (2014) propose a hybrid forecasting model in which the training data is first preprocessed by compressed sensing denoising and then it is used for training certain machine learning techniques including NN and support vector regression (SVR). Ghaffari and Zare (2009) propose an adaptive network-based fuzzy inference system (ANFIS) model for forecasting daily oil price which is first preprocessed with a data filtering algorithm. Chiroma et al. (2015) proposes a hybrid crude oil price forecasting model in which the mataparameters of NN are selected by GA. There is a special kind of hybrid models that is actually ensemble learning paradigm in which the prediction is the integration of the output of several heterogeneous or homogenous models. For instance, Gabralla et al. (2013) proposes an ensemble oil price forecasting model based on SVR, instance based learning (IBL) and K star, and the prediction is generated by taking an average of all the individual forecasts of these machine learning techniques. There are also ensemble models that first decomposed oil price series into several components and then combine the forecasts of each components generated by NNs (Jammazi and Aloui, 2012; Xiong et al., 2013; Yu et al., 2008b, 2016). Essentially, a hybrid learning paradigm consists of two parts. One is a core machine learning technique that is used for training and forecasting the oil price. The other is an additional technique that is used for enhancing the forecasting ability of the entire model (i.e., improve the generalization ability of the core machine learning technique or break the forecasting task into simpler ones).

So far, the machine learning techniques in the above mentioned forecasting models are shallow architectures (e.g. NN with only one hidden layer). Bengio (2009) points out that the functions cannot be efficiently represented (in terms of number of tunable elements) by architectures that are too shallow. In terms of oil price forecasting, shallow architecture base forecasting approaches may fail to model the complex patterns and volatile behaviors of oil price driven by numerous factors. Therefore, a natural idea is to model the oil price with deep architecture based approaches. A deep architecture is the composition of multiple levels of non-linear operations. Recently, deep learning (DL) is becoming a mainstream of machine learning technique, and it has shown strong capacity in various nonlinear modeling tasks such as classification and feature extraction. Hinton and Salakhutdinov (2006) propose a greedy layer wise training strategy which solves the training problem in deep neural network (DNN). After that, relevant algorithms and applications are becoming flourishing. It has turned out to be very good at discovering intricate structures in high-dimensional data and is therefore applicable to many domains of science, business and government (LeCun et al., 2015). It is worth mentioning that, to the best of our knowledge, no DL approach has ever been applied in oil price forecasting. Thus in this work, a deep learning ensemble (DLE) approach for forecasting oil price is proposed by applying one of the most popular DNN model named stacked denoising autoencoders (SDAE) (Vincent et al., 2010) as base model. In the DLE approach, the ensemble method named bootstrap aggregation (bagging) (Breiman, 1996a) is used to generate multiple data set for training SDAEs. Then the trained SDAEs are used to generate predictions. Finally, the predictions of base models are averaged to form the final prediction. For multivariate forecasting, a large set of economic series (e.g., oil production and consumption, stock levels, macroeconomic and financial indicators) are considered as exogenous variables. All these series have ever been used in previous studies (Godarzi et al., 2014; Naser, 2016; Ye et al., 2006; Zagaglia, 2010). In general, a multivariate deep learning ensemble approach is proposed for crude oil price forecasting. Two main contributions of this paper are presented as follows. (1) For the first time, a deep learning approach is introduced to mine the complex relationship of oil price with exogenous variables. (2) A novel deep learning ensemble forecasting approach is built for forecasting oil price series.

The remaining part of this paper is organized as follows. Section 2 describes the formulation of our approach. Section 3 reports the experimental results. Section 4 concludes the paper and outlines the future research direction.

2. Methodology formulation

In this section, a novel deep learning ensemble approach named SDAE bagging (SDAE-B) is formulated for crude oil price forecasting. Deep learning approach (SDAE) and ensemble learning approach (bagging) are respectively introduced in Sections 2.1 and 2.2. Multivariate forecasting approach is described in Section 2.3. Finally, the overall process of the proposed approach is presented in Section 2.4.

2.1. Stacked denoising autoencoders

SDAE (Vincent et al., 2008, 2010) is a popular DNN model that has been proved to have higher prediction accuracy than some competing machine learning models such as SVM, stacked autoencoders (SAE) and deep belief networks (DBN) in a range of classification problems. SDAE is built by stacking several denoising autoencoders (DAEs) which is a special kind of neural network structure. To illustrate the SDAE, autoencoder (AE) and DAE are first introduced.

AE is a one hidden layer neural network where its input and output size are equal. It first maps an input vector $\mathbf{x} \in [0,1]^d$ to a hidden representation $\mathbf{y} \in [0,1]^{d'}$ through a deterministic function:

$$\mathbf{y} = f_\theta(\mathbf{x}) = \phi_f(\mathbf{W}\mathbf{x} + \mathbf{b}) \tag{1}$$

where $f(\mathbf{x})$ is parameterized by $\theta = \{\mathbf{W}, \mathbf{b}\}$, $\mathbf{W}$ is a $d' \times d$ weight matrix, $\mathbf{b}$ is a bias vector, and $\phi_f(\cdot)$ is a nonlinear activation function. Then, the representation $\mathbf{y}$ is mapped back to vector $\mathbf{z} \in [0,1]^d$ in input space:

$$\mathbf{z} = g_{\theta'}(\mathbf{y}) = \phi_g(\mathbf{W}'\mathbf{y} + \mathbf{b}') \tag{2}$$

with $\theta' = \{\mathbf{W}', \mathbf{b}'\}$. Each training $\mathbf{x}^{(i)}$ is thus mapped to a corresponding $\mathbf{y}^{(i)}$ and a reconstruction $\mathbf{z}^{(i)}$. The parameter of this model is optimized to minimize the average reconstruction error.

$$\theta^*, \theta'^* = \arg\min_{\theta,\theta'} \frac{1}{n}\sum_{i=1}^{n} L\left(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}\right) = \arg\min_{\theta,\theta'} \frac{1}{n}\sum_{i=1}^{n} L\left(\mathbf{x}^{(i)}, g_\theta\left(f_\theta\left(\mathbf{x}^{(i)}\right)\right)\right) \tag{3}$$

where L is loss function that can be either traditional squared error $L(\mathbf{x},\mathbf{z}) = \|\mathbf{x}-\mathbf{z}\|^2$ or reconstruction cross-entropy $L_H(x,z) = H(\mathbf{B}_\mathbf{x}\|\mathbf{B}_\mathbf{z}) = -\sum_{k=1}^{d}[\mathbf{x}_k \log \mathbf{z}_k + (1-\mathbf{x}_k)\log(1-\mathbf{z}_k)]$.

Vincent et al. (2010) point out that training a common autoencoder (AE) is unable to guarantee the extraction of useful features just by minimizing the loss function, while DAE can change the reconstruction criteria by denoising (i.e., cleaning partially corrupted input through AE) so as to learn a good representation. The core idea of DAE is to reconstruct a clean input from a corrupted version. Firstly, corrupt the initial input $\mathbf{x}$ into $\tilde{\mathbf{x}}$ by stochastic mapping $\tilde{x} \sim q_D(\tilde{\mathbf{x}}|\mathbf{x})$. Then the corrupted input $\tilde{\mathbf{x}}$ is mapped to a hidden representation $\mathbf{y} = f_\theta(\tilde{\mathbf{x}}) = \phi(\mathbf{W}\tilde{\mathbf{x}} + \mathbf{b})$. Finally, $\mathbf{y}$ is reconstructed to $\mathbf{z} = g_{\theta'}(\mathbf{y})$. For a training set, the best parameters θ and θ' are trained by minimizing the average reconstruction error between $\mathbf{z}$ and the uncorrupted input $\mathbf{x}$. The procedure is shown in Fig. 1.

The deep neural networks SDAE is built by stacking several DAEs, in the same way as stacking Restricted Boltzmann Machines (RBMs) in deep belief networks (Hinton et al., 2006; Hinton and Salakhutdinov, 2006). The procedure is depicted in Fig. 2. It follows the greedy layer wise pretraining procedure (Hinton and Salakhutdinov, 2006) of learning one layer of features at a time. Specifically, the first DAE is trained independently with the training set as its input and the mapping function $f_\theta^{(1)}$ is thus learnt. Then the second DAE is trained with the hidden representation y of the first DAE as its input and the mapping function f_θ

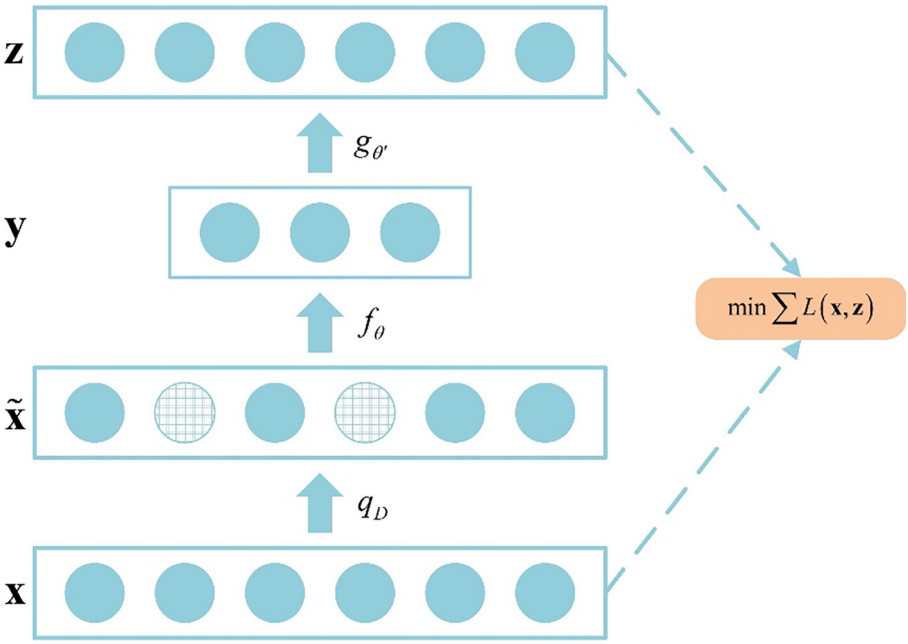

Fig. 1. Denoising autoencoder.

[2] is learnt. All DAEs can be trained independently following the same procedure. A stand-alone supervised learning algorithm (e.g., FNN) is added on top of the structure using hidden representation of the last DAE as its input. Finally, the SDAE is built. Further, the parameters of all layers of SDAE are simultaneously fine-tuned by training algorithms such as gradient descent.

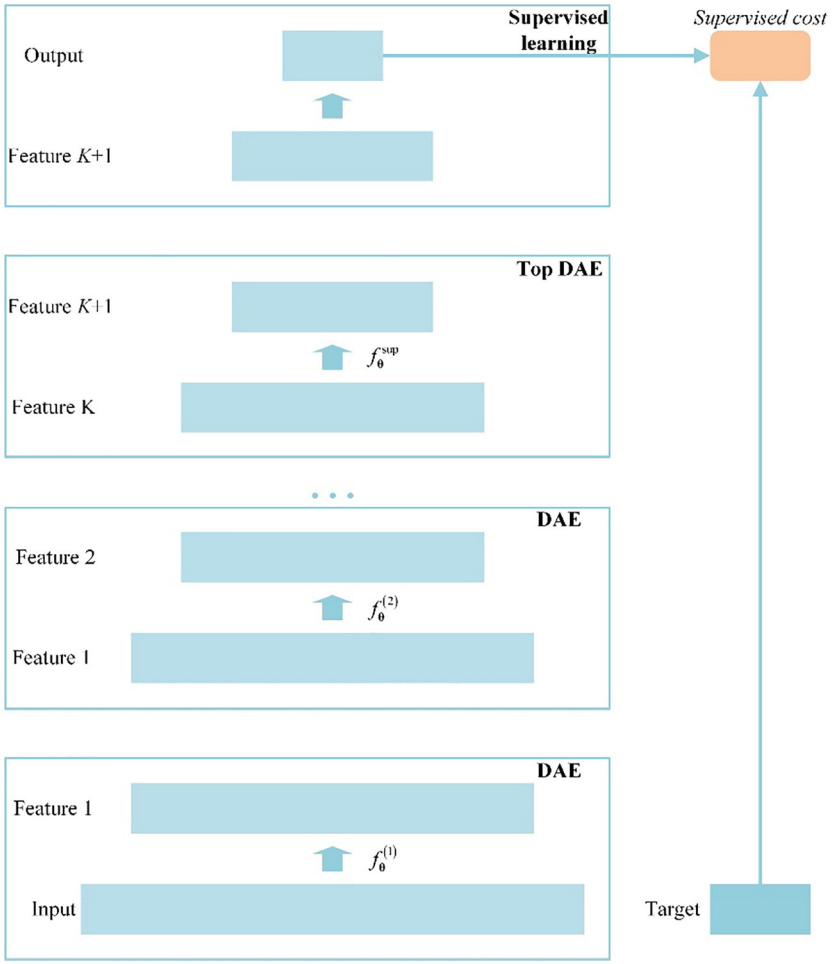

Fig. 2. Stacking denoising autoencoders.

2.2. Bagging

Bagging (Breiman, 1996a) is a powerful ensemble learning algorithm which combines the strengths of a set of base models. The base models are trained with the data generated by bootstrap (Efron and Tibshirani, 1993). Formally, two steps are included to grow an ensemble using bagging. Firstly, generate K bootstrap replicas by randomly selecting M observations with replacement from the training datasets of size M. Secondly, train K models respectively using the bootstrap samples. For prediction, take an average over predictions from all the trained base models.

Bagging generates diverse base models only if the base learning algorithm is unstable, and it can be viewed as ways of exploiting this instability to improve prediction accuracy. Neural network models are instable due to their random initialization process of weights, and bagging neural networks proves to be quite powerful. It has been successfully applied in finance and economics (Kim and Kang, 2010; Kourentzes et al., 2014; Yu et al., 2008a). Based on above reasons, bagging is used to construct an ensemble using SDAEs as base models.

2.3. Multivariate forecasting

Different from time series model, a multivariate model considers not only the autoregressive effect of target series, but also the effect of the exogenous variable to the target series. It can be formalized as a function that models the relationship of dependent and independent variables:

$$y(t+h) = f(s(y), s(x_1), \ldots, s(x_C)) \tag{4}$$

where $y(t+h)$ is value of dependent variable at time $t+h$, and $s(x) = x(t), x(t-1), \ldots, x(t-l_x+1)$ is a set of past values of exogenous variable x with a total number of l_x. Thus the input size is $m = \sum l_y + l_{x_1} + l_{x_2} + \ldots + l_{x_c}$.

2.4. Overall process of SDAE-B approach

The novel SDAE-B approach for forecasting crude oil price is formulated following the five steps below. Fig. 3 illustrates the flowchart of the overall process.

- Data preprocessing: transform and divide the multiple series data into training samples and testing samples.
- Generate multiple training sets: generate k set of replicas of the training samples by bootstrapping.
- Train: train k SDAE models with each set of training samples respectively.
- Forecast: generate k predictions using the k trained SDAE models
- Results aggregation: take the mean value of the k predictions as the final results.

3. Experimental study

In this section, the forecasting ability of SDAE-B is tested against that of some benchmark models. Firstly, data description is given in Section 3.1. Secondly, prediction performance evaluation criteria and statistic test for comparing predictive accuracy are given in Section 3.2. Thirdly, benchmarks and parameter settings are introduced in Section 3.3. Finally the results and analysis are given in Section 3.4.

3.1. Datasets

In this study, West Texas Intermediate (WTI) crude oil spot price series are investigated for forecasting purposes. The original WTI series and its first differenced series are plotted in Fig. 4. Following Zagaglia (2010) and Naser (2016), 198 series including price series, flow and stock series, and macroeconomic and financial series are selected as the exogenous variables. Price series include refiner price of crude oil products and cost of crude oil imports from different regions. Stock and flow series reflect the impact on fundamental supply and demand of oil. They include crude oil production, oil product consumption, and information on rigs and development wells drilled. Macroeconomic and financial series include financial indicators such as indexes of industry, stock and future market, gold price and dollar index. All series are monthly data covering the period from January 1986 to May 2016 including 365 observations.

We select these variables to model the oil price series based on the reasons below. Firstly, they are closely related with the oil prices, and they represent different driving forces of oil price. Secondly, the relationship between oil price series and these factors are noisy, volatile and nonlinear, however, any of them is likely to provide useful information on the movements of oil price at some time point. Thus we can extract more information by including variables as much as possible. Finally, the key reason is because SDAE as a deep neural network model is particularly powerful in modeling high dimensional data. By using all these variables, an ideal circumstance is thus created for modeling the oil price series: on one hand, the merits of SDAE is fully utilized; on the other hand, relatively complete information on oil price movements are provided by using numerous variables.

For modeling and verification purposes, the data are partitioned into two parts. The training samples consist of the first 80% observations of all series and the rest are remained as testing samples

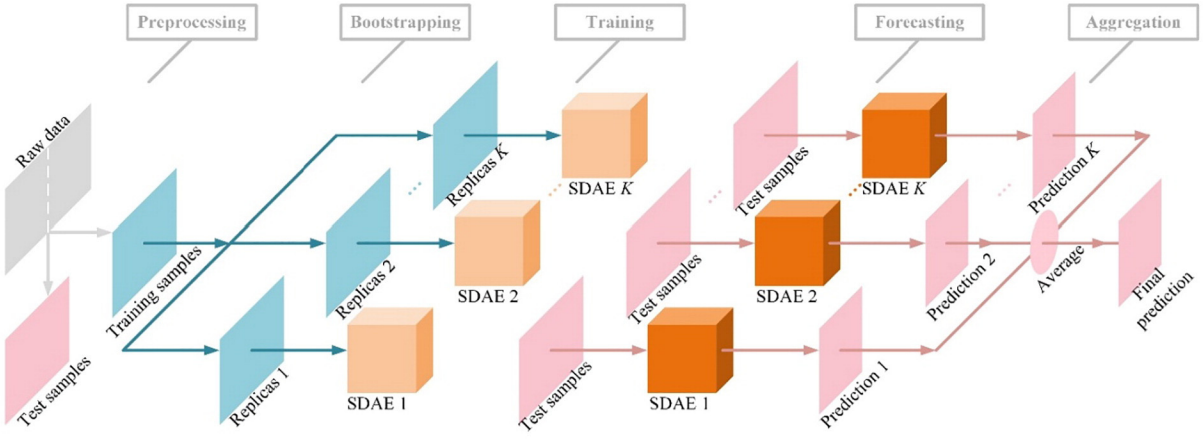

Fig. 3. Flowchart of SDAE-B model.

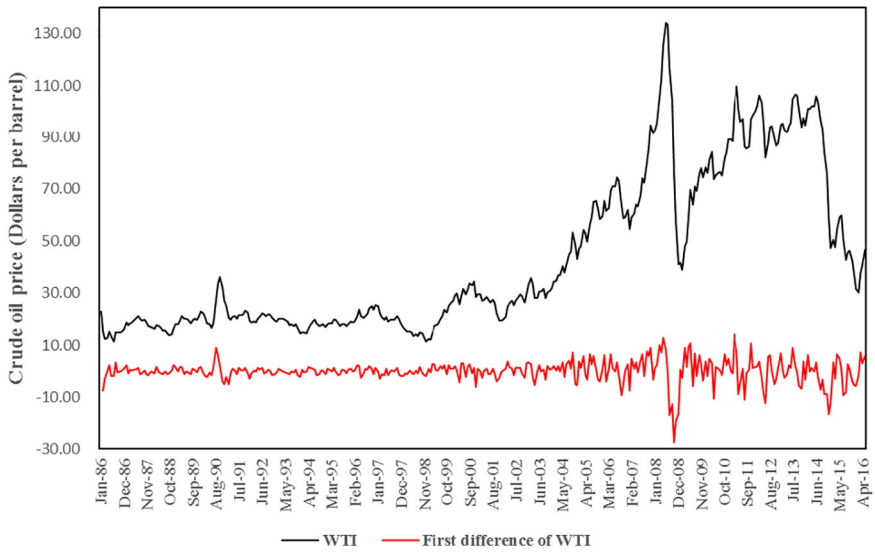

Fig. 4. Monthly WTI crude oil price.

according to Yu et al. (2008b). All the data are obtained from websites including Energy Information Administration (EIA), Federal Reserve Bank (FRB) and yahoo finance. Information of all data are provided in Appendix A.

For model training, oil prices and exogenous variables are preprocessed by means of first difference and normalization. First difference is used to eliminate non-stationarities as suggested by Naser (2016), for the reason that using stationary variables will avoid estimation issues such as parameter explosion (Koop and Korobilis, 2012). All the Phillips-Perron (PP) tests (Peter and Perron, 1988) for the differenced series reject the unit-root null hypothesis, proving that the processed series are stationary. It's worth mentioning that, all the approaches are built for modeling price series rather than return series of oil price.

3.2. Performance evaluation criteria and statistic test

To evaluate the forecasting performance of models from different aspects (i.e., directional prediction and level prediction), three indicators including directional accuracy (DA), mean absolute percentage error (MAPE), root mean square error (RMSE) which have been frequently utilized in recent years (Chiroma et al., 2015; Drachal, 2016; Mostafa and El-Masry, 2016; Yu et al., 2008b, 2014, 2016) are selected:

$$\mathrm{DA} = \frac{1}{N}\sum_{t=1}^{N} a(t) \times 100\% \tag{5}$$

$$\mathrm{MAPE} = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{y(t)-\hat{y}(t)}{y(t)}\right| \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(y(t)-\hat{y}(t)\right)^2} \tag{7}$$

where $y(t)$ and $\hat{y}(t)$ stands for the actual value and predicted value respectively, $a(t)=1$ if $(y(t+1)-y(t))(\hat{y}(t+1)-y(t))\geq 0$ or $a(t)=0$ otherwise, and N is the size of predictions.

To provide statistical evidence of the forecasting ability of proposed model, three tests including Wilcoxon signed rank test (WSRT), forecast encompassing test (FET) and reality check (RC) are performed. WSRT is the best known and most widely used nonparametric inference test (Gibbons and Chakraborti, 2011). The null hypothesis of WSRT is that the loss differential series $d(t)=g(e_A(t))-g(e_B(t))$ has zero median (Diebold and Mariano, 1995), where $e_A(t)$ and $e_B(t)$ are the forecast error series of model A and model B respectively, and $g(\cdot)$ is a loss function (e.g., mean square error). For FET, the Harvey, Leybourne and Newbold (HLN) test (Harvey et al., 1998) is used. The null hypothesis is that the forecast of model A encompasses that of model B (i.e., all information in model B is contained in model A). For RC, test for superior prediction ability (TSPA) (Hansen, 2005) is performed. It's based on White's reality check test (White, 2000) and is more powerful and less sensitive to poor and irrelevant alternatives. The null hypothesis of SPA test is that the predictive performance of model A is no better than that of model B.

3.3. Benchmarks and parameter settings

To test the superiority of the proposed model, seven forecasting models are built and used as benchmarks. Firstly, random walk (RW) model and Markov regime switching (MRS) model are selected. Then, four models include two popular machine learning models (i.e., feedforward neural network (FNN) and SVR) and their ensemble form (i.e., FNN-B and SVR-B). Their ensemble strategy is bagging which is consistent with the ensemble strategy used in our proposed model. The last benchmark model is SDAE. The reasons of selecting the above benchmark models are as follows. For RW, it is a very popular benchmark for oil price according to Alquist et al. (2013). For MRS, it has noticeable impact in finance as one of the complex, non-linear models proposed in the econometrics literature (Brooks, 2014), and has shown its capability in modeling oil prices (Naifar and Al Dohaiman, 2013; Wang et al., 2016; Zhang and Zhang, 2015). For FNN and SVR, they are the most widely used machine learning models in crude oil price forecasting as introduced in Section 1, and their ensemble forms FNN-B and SVR-B are also proved to be powerful and popular forecasting techniques (Kim and Kang, 2010; Yu et al., 2008a). For SDAE, it is to test the capability of bagging algorithm in the proposed model.

The MRS model is built following Naifar and Al Dohaiman (2013), and a Markov two-regime switching model is specified. For machine learning based models, the FNN model is a standard two layer neural network model including a hidden layer and an output layer. The number of nodes in the hidden layer is set to 20 as Godarzi et al. (2014) note that small number of hidden neurons results in inaccuracy of the correlation between inputs and outputs while too large number of hidden neurons results in local optimums. Hastie et al. (2009) concluded that

the typical number of nodes is in the range of 5 to 100, and it's unnecessary to use cross validation. In the SVR model, the Gaussian kernel function is selected (Yu et al., 2014). To be consistent with FNN, the penalty coefficient and kernel scale of SVR are not cross validated. They are respectively set as $iqr(Y)/1.349$ and 1, where $iqr(Y)$is the interquartile range of processed target series. The SDAE model is a three hidden layer neural network model by stacking 2 DAEs with a supervised FNN on top of the architecture. The numbers of nodes in the first to third hidden layers are set to 200, 100 and 10 according to the size of feature space. The log-sigmoid transfer function is used in all the hidden and output layers as suggested by Ng (2011). The number of epochs for unsupervised pretraining of DAEs, supervised pretraining of FNN and global fine tuning are all set to 500 based on trial and error method. The noise type in DAEs is additive Gaussian noise $\tilde{\mathbf{x}}|\mathbf{x} \sim N(\mathbf{x}, \sigma^2 I)$ with a level of 0.2 based on Vincent et al. (2010). For ensemble models (i.e., FNN-B, SVR-B and proposed SDAE-B), the numbers of ensemble members are all set to 100. The choice of number of ensemble member is actually a trade-off between computational complexity and accuracy. However, the forecast error of an ensemble converges to a certain level quickly as the ensemble member grows, thus there is no need to choose an excessive large number. The lag order for all the variables and oil price in multivariate forecasting models are set to 1, i.e., to learn a function $f(\cdot)$ satisfying $y(t+1)=f(y(t),x_1(t),\ldots,x_c(t))$. By this means, a short term nonlinear dependency can be learned between input/output data.

All models are running 10 times using Matlab R2016a software on a server with 16 Core CPU of i3 3.30 GHz, RAM size of 16 GB.

3.4. The superiority of SDAE-B

In this section, the evaluation results of seven benchmark models and the proposed model are given respectively based on the out of sample predictions. Then three tests are conducted in order to test the significance level of the difference between predictions of a pair of models.

The actual values of the crude oil prices and the forecasts of each model are shown in Fig. 5(a), and the squared forecast errors of each model are shown in Fig. 5(b).

Table 1
Prediction accuracy of all models.

Competing models	DA	MAPE	RMSE	Elapsed time (s)
RW	0.493 (0.036)	0.074 (0.005)	6.832 (0.330)	0.001
MRS	0.542 (0.000)	0.060 (0.000)	5.616 (0.000)	11.881
SVR	0.556 (0.000)	0.063 (0.000)	5.867 (0.000)	0.024
SVR-B	0.556 (0.000)	0.064 (0.000)	5.872 (0.002)	0.194
FNN	0.603 (0.032)	0.058 (0.002)	5.428 (0.215)	1.047
FNN-B	0.647 (0.019)	0.054 (0.000)	5.079 (0.018)	7.335
SDAE	0.651 (0.028)	0.053 (0.002)	5.047 (0.138)	10.562
SDAE-B	**0.672** (0.020)	**0.053** (0.000)	**4.995** (0.020)	155.898

The value in boldface represents the best performance amongst 8 models in terms of DA, MAPE and RMSE.

Table 1 shows the forecasting performance of different models. From second to the fourth columns, the mean value and the standard deviation (in the bracket) of the statistics are listed. The last column reports the average running time of each model.

Firstly, comparing the five single models including RW, MRS, FNN, SVR and SDAE, it can be easily found that the deep learning model SDAE achieves the highest DA, the lowest MAPE and RMSE among the single models. The possible reason is as follows. Taking the input of a prediction model as features, the model is actually learning a representation that maps the features and the target variable. As Bengio (2013) noted, a weakness exists in many traditional shallow learning algorithms when dealing with data of various features: the inability to extract and organize the discriminative information from the data. Since modeling the relationship between oil price and so many variables can be a difficult task, shallow models may fail to learn a useful representation, and generate inferior out of sample forecast as a result. On the contrary, SDAE has a deep architecture. It can learn useful information from features by abstracting multiple levels of representation, and thus leads to superior predictive ability with out of sample data. For the rest of the single models, it is seen that MRS performs better than SVR and RW in terms of MAPE and RMSE, and worse than FNN under all criteria. MRS shows its capacity as a nonlinear econometric model.

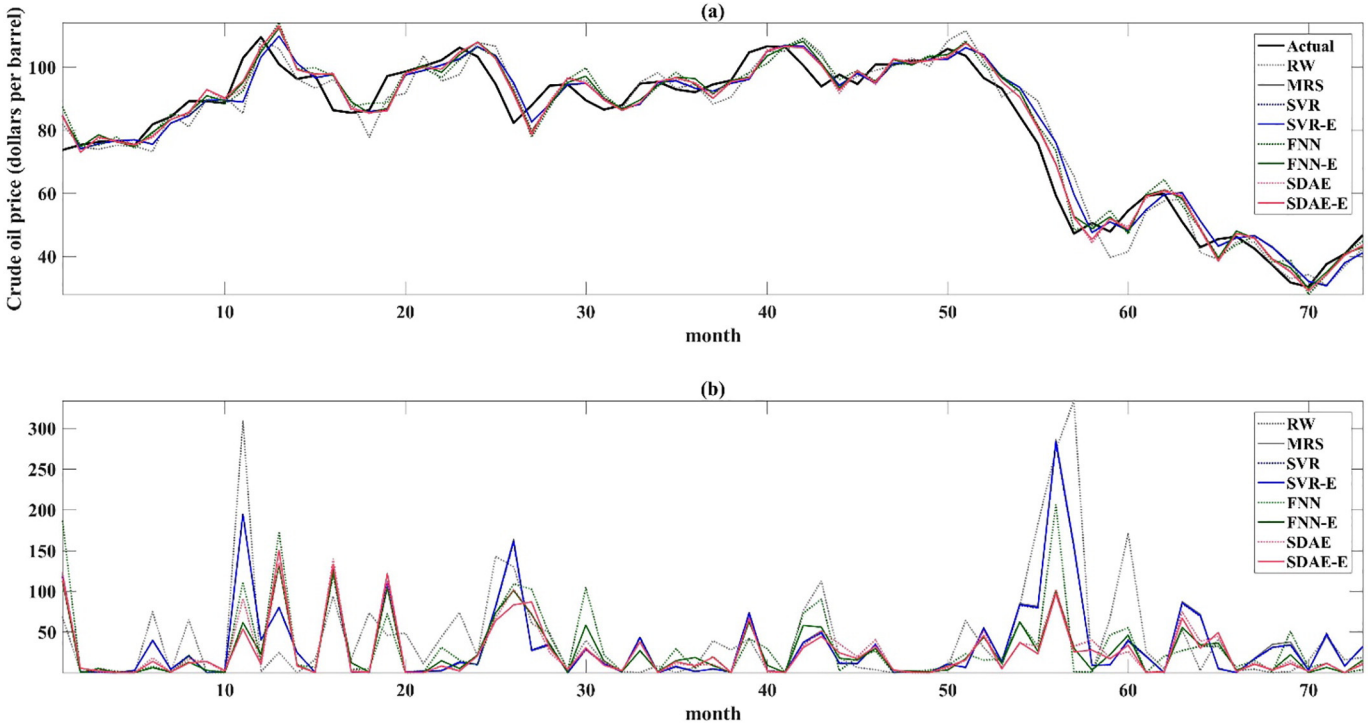

Fig. 5. Actual value and forecasts: (a) predicted series, (b) squared forecast error.

Table 2
Wilcoxon signed rank test.

	RW	MRS	SVR	SVR-B	FNN	FNN-B	SDAE	SDAE-B
RW		0.082	0.102	0.101	0.035	0.037	0.033	0.025
MRS	0.082		0.332	0.252	0.694	0.074	0.071	0.027
SVR	0.102	0.332		0.443	0.654	0.076	0.086	0.033
SVR-B	0.101	0.252	0.443		0.670	0.081	0.085	0.033
FNN	0.035	0.694	0.654	0.670		0.017	0.062	0.016
FNN-B	0.037	0.074	0.076	0.081	0.017		0.443	0.326
SDAE	0.033	0.071	0.086	0.085	0.062	0.443		0.051
SDAE-B	0.025	0.027	0.033	0.033	0.016	0.326	0.051	

Table 4
Reality check.

	RW	MRS	SVR	SVR-B	FNN	FNN-B	SDAE	SDAE-B
RW		0.000	0.000	0.000	0.000	0.000	0.000	0.000
MRS	0.990		0.930	0.935	0.880	0.000	0.005	0.000
SVR	0.980	0.045		0.855	0.020	0.000	0.000	0.000
SVR-B	1.000	0.025	0.130		0.020	0.000	0.000	0.000
FNN	1.000	0.180	0.930	0.955		0.000	0.000	0.000
FNN-B	1.000	1.000	1.000	0.995	1.000		0.915	0.110
SDAE	0.995	0.995	0.995	0.985	1.000	0.155		0.055
SDAE-B	1.000	1.000	0.995	1.000	1.000	0.915	0.920	

As for the naïve RW, its prediction is guided by stochastic behavior, thus it achieves the lowest forecasting performance with no doubt.

When comparing single machine learning model with its corresponding ensemble model, it is found that FNN-B and SDAE-B perform better than single FNN and SDAE under all criteria. Moreover, the performance of these two ensemble models is more stable as their standard deviations are much smaller. The results indicate the effectiveness of bagging. The reason is as follows. On one hand, individual models are overfit as they overrely on sample data and underestimate variance (Grushka-Cockayne et al., 2017). However, combining forecasts (i.e., ensemble modeling or model averaging) can reduce overfitting or generalization error as noted by Goodfellow et al. (2016). One the other hand, bagging works well for unstable models and neural network models are unstable due to the random initialization of weights of nodes (Breiman, 1996b). On the contrary, SVR-B performs worse than the base model SVR in terms of MAPE and RMSE. The possible reason is that SVR is stable as it is seen that the standard deviations of the three indicators of SVR are very small. However, the vital element of substantial gains in accuracy by bagging is the instability of the prediction method (Breiman, 1996a).

Additionally, the last column of Table 1 shows that the proposed model has the largest computational cost as the average running time of a SDAE-B model is about 156 s. However, the computational cost can barely be a priority, as the computational capacity of hardware is fast growing and technique such as parallel computing can be utilized.

Finally, the results show that SDAE-B has the best prediction performance among all the models as it has the highest DA, lowest MAPE and lowest RMSE. It indicates that the proposed SDAE-B approach is powerful for oil price forecasting.

Tables 2–4 respectively report the empirical results of Wilcoxon signed rank test, forecast encompassing test and test for reality check. In these tables the p values of relevant statistics between pairs of models are listed. For illustration purpose, in Table 2, the p values in row 2, column 3 is 0.082, thus the test rejects the null hypothesis (i.e., there is no significant differences between the forecasts of RW and MRS) under the confidence level of 90%. For forecast encompassing test in Table 3, the p values in row 2, column 3 is 0.146, thus the test fails to reject the null hypothesis (i.e., the forecasts of MRS encompasses that of RW). However, the p values in row 3, column 2 is 0.000, thus the test rejects the null hypothesis (i.e., the forecasts of RW encompasses that of MRS) under the confidence level of 99%. For reality check in Table 4, the p values in row 2, column 3 is 0.000, thus the test rejects the null hypothesis (i.e., the forecasts of MRS is no better than that of RW) under the confidence level of 99%. On the contrary, the p values in row 3, column 2 is 0.990, the test fails to reject the null hypothesis (i.e., the forecasts of RW is no better than that of MRS).

Table 3
Forecast encompassing test.

	RW	MRS	SVR	SVR-B	FNN	FNN-B	SDAE	SDAE-B
RW		0.146	0.198	0.190	0.059	0.503	0.468	0.506
MRS	0.000		0.013	0.012	0.005	0.176	0.445	0.597
SVR	0.000	0.137		0.108	0.091	0.860	0.858	0.869
SVR-B	0.000	0.134	0.877		0.095	0.867	0.863	0.873
FNN	0.001	0.012	0.018	0.018		0.885	0.601	0.809
FNN-B	0.000	0.002	0.003	0.003	0.000		0.089	0.466
SDAE	0.000	0.003	0.003	0.003	0.001	0.279		0.940
SDAE-B	0.000	0.003	0.002	0.002	0.000	0.041	0.020	

Focusing on Wilcoxon signed rank test in Table 2, in the last column, all the p values are smaller than 0.1 except for that in row 7 column 9, indicating there are significant differences between the forecasts of proposed SDAE-B with that of six competing models (i.e., RW, MRS, SVR, SVR-B, FNN and SDAE), under the confidence level of 90%. Considering the prediction results from Table 1, it can be concluded that the prediction accuracy of proposed SDAE-B is better than that of the above six models from statistic point of view. As for the exceptional case, the p value between FNN-B and SDAE-B is 0.326, thus the test fails to reject the null hypothesis of no differences between these two models. The possible reason is that bagging as a powerful ensemble strategy greatly improves the prediction accuracy of single FNN model. Specifically, the p values of FNN with FNN-B are 0.017, and that of SDAE with SDAE-B is 0.051. It indicates that there are significant differences between single model and its corresponding ensemble model. It again confirms the strong capability of bagging.

Focusing on forecast encompassing test in Table 3, the p values in the last column show that all the tests fail to reject the null hypothesis that the forecasts of SDAE-B encompass that of a competing model. On the contrary, the p values in the last row show that all the tests reject the null hypothesis that the forecasts of a competing model encompass that of the SDAE-B under the confidence level of 95%. Thus it can be concluded that SDAE-B is a more efficient approach of which more information is contained in the forecasts.

With regard to reality check in Table 4, similar conclusions can be drawn. Except for the case in row 7, column 9, the p values in the last column are all smaller than 0.1, indicating that these tests reject the null hypothesis that performance of SDAE-B is no better than that of a competing model. Accordingly, p values from the last row are close to 1, thus these tests fail to reject the null hypothesis that the performance of a competing model is no better than that of the SDAE-B.

In general, according to the prediction performance and statistic tests, the predictive ability of the proposed SDAE-B approach is verified with statistical evidence.

3.5. Summarization

Based on the experiments presented in Section 3.4, five conclusions can be drawn. (1) SDAE performs better than the competing single models. (2) Bagging as an ensemble learning approach is effective in enhancing the prediction accuracy and stability of single models. (3) The superiority of the proposed model is verified by statistic tests. (4) The proposed SDAE-B model is a promising approach in crude oil price forecasting.

4. Conclusion

Due to the complex relationship of crude oil price with various factors, a deep learning ensemble approach named SDAE-B is proposed

for crude oil price forecasting. In this approach, deep learning model SDAE is applied to learn useful representations from data and to generate forecasts. Bagging as a powerful ensemble method combines the strength of multiple SDAEs and thus generates an ensemble model of better performance. In the empirical study, the proposed SDAE-B outperforms benchmark models including econometric models (i.e., RW and MRS), machine learning models of shallow architectures (i.e., SVR, SVR-B, FNN and FNN-B) and its base model SDAE. Statistic tests further confirm the superiority of the proposed model, indicating the proposed approach can be used as a promising forecasting tool for crude oil price.

Despite the capability of the proposed deep learning ensemble approach, there is still room for improvement. It is well known that irregular factors such as extreme climate, politic risks, and psychological factor also have great impact on oil price volatility, yet it's rather challenging to quantify the effect of them. We believe that better predictive accuracy can be generated by quantifying these factors and utilize the information from them. We will look into this issue in the future research.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.eneco.2017.05.023.

Acknowledgements

We are grateful to the editor and two anonymous referees for offering valuable suggestions that greatly improved the presentation of this article.

This work was supported by the National Natural Science Foundation of China under Grant 71425002 and 71433001.

References

Alquist, R., Kilian, L., Vigfusson, R.J., 2013. Chapter 8 - forecasting the price of oil. In: Graham, E., Allan, T. (Eds.), Handbook of Economic Forecasting. Elsevier, pp. 427–507.
Baumeister, C., Kilian, L., 2016. Forty years of oil price fluctuations: why the price of oil may still surprise us. J. Econ. Perspect. 30, 139–160.
Bengio, Y., 2009. Learning deep architectures for AI. Found. Trends Mach. Learn. 2, 1–127.
Bengio, Y., 2013. Deep learning of representations: looking forward. In: Dediu, A.-H., Martín-Vide, C., Mitkov, R., Truthe, B. (Eds.), Statistical Language and Speech Processing: First International Conference, SLSP 2013, Tarragona, Spain, July 29–31, 2013. Proceedings. Springer, Berlin Heidelberg, pp. 1–37.
Breiman, L., 1996a. Bagging predictors. Mach. Learn. 24, 123–140.
Breiman, L., 1996b. Heuristics of Instability and Stabilization in Model Selection. pp. 2350–2383.
Brooks, C., 2014. Introductory Econometrics for Finance. Cambridge University Press.
Chiroma, H., Abdulkareem, S., Herawan, T., 2015. Evolutionary neural network model for West Texas intermediate crude oil price prediction. Appl. Energy 142, 266–273.
Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. J. Bus. Econ. Stat. 13, 253–263.
Drachal, K., 2016. Forecasting spot oil price in a dynamic model averaging framework — have the determinants changed over time? Energy Econ. 60, 35–46.
Efron, B., Tibshirani, R., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.
Gabralla, L.A., Jammazi, R., Abraham, A., 2013. Oil price prediction using ensemble machine learning. Computing, Electrical and Electronics Engineering (ICCEEE), 2013 International Conference on, pp. 674–679.
Ghaffari, A., Zare, S., 2009. A novel algorithm for prediction of crude oil price variation based on soft computing. Energy Econ. 31, 531–536.
Gibbons, J.D., Chakraborti, S., 2011. Nonparametric statistical inference. In: Lovric, M. (Ed.), International Encyclopedia of Statistical Science. Springer, Berlin Heidelberg, pp. 977–979.
Godarzi, A.A., Amiri, R.M., Talaei, A., Jamasb, T., 2014. Predicting oil price movements: a dynamic artificial neural network approach. Energ Policy 68, 371–382.
Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning.
Grushka-Cockayne, Y., Jose, V.R.R., Lichtendahl Jr., K.C., 2017. Ensembles of overfit and overconfident forecasts. Manag. Sci. 63, 1110–1130.
Hansen, P.R., 2005. A test for superior predictive ability. J. Bus. Econ. Stat. 23, 365–380.
Harvey, D.S., Leybourne, S.J., Newbold, P., 1998. Tests for forecast encompassing. J. Bus. Econ. Stat. 16, 254–259.
Hastie, T., Tibshirani, R., Friedman, J., 2009. Neural Networks, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer New York, New York, NY, pp. 389–416.
Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science 313, 504–507.
Hinton, G.E., Osindero, S., Teh, Y.-W., 2006. A fast learning algorithm for deep belief nets. Neural Comput. 18, 1527–1554.
Jammazi, R., Aloui, C., 2012. Crude oil price forecasting: experimental evidence from wavelet decomposition and neural network modeling. Energy Econ. 34, 828–841.
Kaboudan, M.A., 2001. Compumetric forecasting of crude oil prices. IEEE C Evol. Comput. 283–287.
Kim, M.-J., Kang, D.-K., 2010. Ensemble with neural networks for bankruptcy prediction. Expert Syst. Appl. 37, 3373–3379.
Koop, G., Korobilis, D., 2012. Forecasting inflation using dynamic model averaging*. Int. Econ. Rev. 53, 867–886.
Kourentzes, N., Barrow, D.K., Crone, S.F., 2014. Neural network ensemble operators for time series forecasting. Expert Syst. Appl. 41, 4235–4244.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.
Moshiri, S., Foroutan, F., 2006. Forecasting nonlinear crude oil futures prices. Energy J. 27, 81–95.
Mostafa, M.M., El-Masry, A.A., 2016. Oil price forecasting using gene expression programming and artificial neural networks. Econ. Model. 54, 40–53.
Naifar, N., Al Dohaiman, M.S., 2013. Nonlinear analysis among crude oil prices, stock markets' return and macroeconomic variables. Int. Rev. Econ. Financ. 27, 416–431.
Naser, H., 2016. Estimating and forecasting the real prices of crude oil: a data rich model using a dynamic model averaging (DMA) approach. Energy Econ. 56, 75–87.
Ng, A., 2011. Sparse autoencoder. CS294A Lecture notes 72, pp. 1–19.
Peter, C.B.P., Perron, P., 1988. Testing for a unit root in time series regression. Biometrika 75, 335–346.
Shin, H., Hou, T., Park, K., Park, C.-K., Choi, S., 2013. Prediction of movement direction in crude oil prices based on semi-supervised learning. Decis. Support. Syst. 55, 348–358.
Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.-A., 2008. Extracting and composing robust features with denoising autoencoders. Proceedings of the 25th International Conference on Machine Learning. ACM, Helsinki, Finland, pp. 1096–1103.
Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A., 2010. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. 11, 3371–3408.
Wang, Y., Wu, C., Yang, L., 2016. Forecasting crude oil market volatility: a Markov switching multifractal volatility approach. Int. J. Forecast. 32, 1–9.
White, H., 2000. A reality check for data snooping. Econometrica 68, 1097–1126.
Xie, W., Yu, L., Xu, S., Wang, S., 2006. A new method for crude oil price forecasting based on support vector machines. In: Alexandrov, V.N., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (Eds.), Computational Science – ICCS 2006: 6th International Conference, Reading, UK, May 28–31, 2006, Proceedings, Part IV. Springer, Berlin Heidelberg, pp. 444–451.
Xiong, T., Bao, Y., Hu, Z., 2013. Beyond one-step-ahead forecasting: evaluation of alternative multi-step-ahead forecasting models for crude oil prices. Energy Econ. 40, 405–415.
Ye, M., Zyren, J., Shore, J., 2006. Forecasting short-run crude oil price using high- and low-inventory variables. Energ Policy 34, 2736–2743.
Yu, L., Wang, S., Lai, K.K., 2008a. Credit risk assessment with a multistage neural network ensemble learning approach. Expert Syst. Appl. 34, 1434–1444.
Yu, L., Wang, S., Lai, K.K., 2008b. Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. Energy Econ. 30, 2623–2635.
Yu, L., Zhao, Y., Tang, L., 2014. A compressed sensing based AI learning paradigm for crude oil price forecasting. Energy Econ. 46, 236–245.
Yu, L., Zhao, Y., Tang, L., 2016. Ensemble forecasting for complex time series using sparse representation and neural networks. J. Forecast. (n/a-n/a).
Zagaglia, P., 2010. Macroeconomic factors and oil futures prices: a data-rich model. Energy Econ. 32, 409–417.
Zhang, Y.-J., Zhang, L., 2015. Interpreting the crude oil price movements: evidence from the Markov regime switching model. Appl. Energy 143, 96–109.