Applied Energy 156 (2015) 251–267

Contents lists available at ScienceDirect

Applied Energy

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

A decomposition–ensemble model with data-characteristic-driven reconstruction for crude oil price forecasting

Lean Yu, Zishu Wang, Ling Tang *

School of Economics and Management, Beijing University of Chemical Technology, Beijing 100029, China

HIGHLIGHTS

- A decomposition–ensemble model is proposed for crude oil price forecasting.
- A data-characteristic-driven reconstruction is formulated and introduced.
- Four steps are involved: decomposition, reconstruction, prediction and ensemble.
- Empirical study statistically verifies the effectiveness of the proposed model.

ARTICLE INFO

Article history:
Received 8 April 2015
Received in revised form 29 June 2015
Accepted 12 July 2015

Keywords:
Decomposition ensemble model
Divide and conquer
Data characteristics
Reconstruction
Time series analysis
Crude oil price forecasting

ABSTRACT

To enhance prediction accuracy and reduce computation complexity, a decomposition–ensemble methodology with data-characteristic-driven reconstruction is proposed for crude oil price forecasting, based on two promising principles of "divide and conquer" and "data-characteristic-driven modeling". Actually, this proposed model improves the existing decomposition–ensemble techniques in the "divide and conquer" framework, by formulating and incorporating a data-characteristic-driven reconstruction method based on the "data-characteristic-driven modeling". Four main steps are involved in the proposed methodology, i.e., data decomposition for simplifying the complex data, component reconstruction based on the "data-characteristic-driven modeling" for capturing inner factors and reducing computational cost, individual prediction for each reconstructed component via a certain artificial intelligence (AI) tool, and ensemble prediction for final output. In the proposed data-characteristic-driven reconstruction, all decomposed modes are thoroughly analyzed to explore the hidden data characteristics, and are accordingly reconstructed into some meaningful components. For illustration and verification, the West Texas Intermediate (WTI) and Brent crude oil spot prices are used as the sample data, and the empirical results indicate that the proposed model statistically outperforms all considered benchmark models (including popular AI single models, typical decomposition–ensemble models without reconstruction, and similar decomposition–ensemble models with other existing reconstruction methods), since it has higher prediction accuracy and less computational time.

1. Introduction

Crude oil price forecasting has become an increasingly hot issue within the research fields of data analysis and prediction, due to the important role in global energy system and even in global economic system. However, crude oil price forecasting has fully been proven to be an extremely difficult task [1]. On the one hand, like other commodities, crude oil price is driven by various market factors, e.g., supply and demand. On the other hand, as a special energy resource, crude oil price is strongly influenced by some exogenous factors, such as irregular events [2], global economic status [3], speculation activities [4], and political and social attitudes [5], whose effects on crude oil market are sometimes hard to quantify. Therefore, this paper focuses on crude oil price forecasting, especially to investigate the inner hidden factors and further to improve model performance in terms of prediction accuracy and time-saving.

According to existing literature, abundant time series forecasting models have been proposed and applied to crude oil price forecasting, which can be generally described as [6]:

$$x_{t+h} = f(X_t) + \varepsilon_t \tag{1}$$

* Corresponding author at: School of Economics and Management, Beijing University of Chemical Technology, 15 Beisanhuan East Road, Beijing 100029, China. Tel./fax: +86 10 6441 2210.

E-mail address: tangling@mail.buct.edu.cn (L. Tang).

http://dx.doi.org/10.1016/j.apenergy.2015.07.025
0306-2619/

Nomenclature

AI	artificial intelligence
ANN	artificial neural network
EEMD	ensemble empirical mode decomposition
EMD	empirical mode decomposition
FNN	feed-forward neural network
FFT	fast Fourier transform
ICSS	iterative cumulative sums of squares
IMF	intrinsic mode function
LSSVR	least squares support vector regression
SVR	support vector regression

where x_t denotes the crude oil price at time t, $X_t = \{x_{t-1}, x_{t-2},\ldots x_{t-l}\}$ are the history values before period t with lag l, h is the prediction horizon, and ε_t is the prediction errors following independent identical distribution. According to different function designs $f(*)$ and the corresponding parameter evaluation methods, the existing models for crude oil price forecasting can fall into three main types: traditional econometric models with relatively simple fixed functions and strict data assumptions (e.g., stationarity and linearity) in parameter evaluation, artificial intelligence (AI) techniques with flexible functions and powerful self-learning capability in model training, and currently popular hybrid models combining several single models systematically.

As for traditional econometric models, auto-regressive integrated moving average (ARIMA), generalized autoregressive conditional heteroskedasticity (GARCH), random walk (RW), vector auto-regression (VAR) and error correction models (ECM) have popularly been used for crude oil price forecasting. For example, Xiang and Zhuang [7] utilized the ARIMA model to predict the Brent monthly crude oil prices for the sample period from November 2012 to April 2013. Nomikos and Andriosopoulos [8] estimated the conditional mean and volatility of West Texas Intermediate (WTI) daily crude oil spot prices from December 9, 2000 to January 2, 2010, based on the GARCH family models. The RW, one basic time series model, can also be employed as a benchmark model to forecast oil price movements [9]. Mirmirani and Li [10] used the VAR model to predict the US monthly oil price covering the period from January 1980 to November 2002. Lanza et al. [11] used the ECM model to predict the WTI and Brent weekly crude oil prices with the sample period from 1994 to 2002.

As for the AI techniques, artificial neural networks (ANN), support vector regression (SVR) and least support vector regression (LSSVR) might be the most predominant models for crude oil price forecasting, and the empirical investigations have repeatedly shown their superiority over the traditional linear models. As for the ANN, Movagharnejad et al. [12] introduced the ANN to forecast the quantitative data of crude oil prices over the period from January 2000 to April 2010. Chiroma et al. [13] presented an evolutionary neural network to predict the WTI monthly crude oil price data from May 1987 to December 2011. As for the SVR, Xie et al. [14] compared the SVR with the ARIMA and back-propagation neural network (BPNN), and witnessed the superiority of the SVR in the prediction for the WTI monthly prices from January 1970 to December 2003. Khashman and Nwulu [15] predicted the WTI weekly spot crude oil prices from January 03, 1986 to December 25, 2009, based on the SVR model. Similarly, Li and Ge [16] predicted the crude oil prices from May 1994 to December 1995 based on a ε-SVR model with dynamic errors correction. As for the LSSVR, Li et al. [17] predicted the WTI weekly data from January 4, 2008 to October 18, 2013, and argued that the LSSVR outperformed the ARIMA, SVR and BPNN models. However, these AI models have their own weaknesses, e.g., parameter sensitiveness and potential over-fitting [18].

As for hybrid models, under the promising concept of "divide and conquer" (or "decomposition and ensemble") [19], a series of decomposition–ensemble learning paradigms have currently been developed and become a predominant type for crude oil price analysis and forecasting. In a typical decomposition–ensemble model, three main steps are included, i.e., data decomposition for simplifying the complex data, individual prediction for each decomposed mode, and ensemble prediction for final prediction result [18,20,21]. Existing studies have demonstrated that decomposition–ensemble models, in the effective framework of "divide and conquer", can provide satisfactory results for both capturing inner factors and enhancing prediction accuracy. Some current studies on the crude oil price forecasting via decomposition–ensemble models can be listed as below. Yu et al. [20] introduced empirical mode decomposition (EMD) to decompose the original data of the WTI daily crude oil spot prices from January 1, 1986 to September 30, 2006 and the Brent daily crude oil spot prices form May 20, 1987 to September 30, 2006, and produced better prediction results. Tang et al. [22] formulated a novel decomposition–ensemble learning paradigm for crude oil price forecasting by utilizing the data decomposition tool of complementary ensemble EMD (CEEMD), and the results supported the efficiency of the decomposition strategy in improving model performance. Yu et al. [1] proposed a compressed sensing based AI learning paradigm for daily crude oil price forecasting and achieved a similar conclusion, with the sample period from January 3, 2011 to July 17, 2013. Yu et al. [23] proposed a novel learning paradigm based on ensemble EMD (EEMD) and extended extreme learning machine (EELM), to predict the WTI daily crude oil prices from January 2, 1986 to October 21, 2013. All above empirical studies showed that the decomposition–ensemble strategy can significantly improve the model performance in crude oil price prediction.

However, even though these decomposition–ensemble models can effectively model the complex data of crude oil price compared with single models, another important issue may arise concerning the model complexity and computational cost. Since decomposition–ensemble models decompose the original data into a series of modes, modeling all decomposed modes might be a quite time-consuming process in the step of individual prediction, even sometimes leading to a poor final result since the estimation errors for all modes can be accumulated in the ensemble prediction step. To address this problem, an additional step of component reconstruction has been introduced into the typical decomposition–ensemble models, between the steps of data decomposition and individual prediction. In component reconstruction, the decomposed modes obtained from the data decomposition step are reconstructed into some certain components for further analysis in the next step of individual prediction.

Accordingly, some modified decomposition–ensemble models with reconstruction have currently been developed and shown effective in understanding inner factors, enhancing prediction accuracy and reducing computational cost. For example, Wang et al. [24] implemented the run-length-judgment method to reconstruct the specific modes decomposed by EMD into high frequency, medium frequency, low frequency and trend sequences, and the empirical analysis showed that the novel decomposition–ensemble