Geoscience Frontiers xxx (2016) 1–5

HOSTED BY ELSEVIER

Contents lists available at ScienceDirect

China University of Geosciences (Beijing)

Geoscience Frontiers

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

Research paper

A new approach for crude oil price prediction based on stream learning

Shuang Gao, Yalin Lei*

School of Humanities and Economic Management, China University of Geosciences, Beijing 100083, China

ARTICLE INFO

Article history:
Received 27 May 2016
Received in revised form
30 July 2016
Accepted 4 August 2016
Available online xxx

Keywords:
Crude oil
Economic geology
Prediction model
Machine learning
Stream learning

ABSTRACT

Crude oil is the world's leading fuel, and its prices have a big impact on the global environment, economy as well as oil exploration and exploitation activities. Oil price forecasts are very useful to industries, governments and individuals. Although many methods have been developed for predicting oil prices, it remains one of the most challenging forecasting problems due to the high volatility of oil prices. In this paper, we propose a novel approach for crude oil price prediction based on a new machine learning paradigm called stream learning. The main advantage of our stream learning approach is that the prediction model can capture the changing pattern of oil prices since the model is continuously updated whenever new oil price data are available, with very small constant overhead. To evaluate the forecasting ability of our stream learning model, we compare it with three other popular oil price prediction models. The experiment results show that our stream learning model achieves the highest accuracy in terms of both mean squared prediction error and directional accuracy ratio over a variety of forecast time horizons.

1. Introduction

Crude oil is a natural liquid fossil fuel found in geological formations beneath the earth's surface. It has mostly been extracted by oil drilling, which comes after the studies of structural geology, sedimentary basin analysis, and reservoir characterization (Guerriero et al., 2012). Crude oil is one of the most important energy resources on earth. So far, it remains the world's leading fuel, with nearly one-third of global energy consumption.

Crude oil prices are determined by many factors and have a big impact on the global environment and economy. Although crude oil prices were firm in early 2014, they fell sharply from mid 2014. In January 2016, the U.S. refiner acquisition cost for crude oil imports, as a proxy for world oil price, is only $28.81 per barrel on average, and the West Texas Intermediate (WTI) crude oil spot price, as the benchmark oil price in North America, is only $31.68 per barrel on average (EIA, 2016). The prices have dropped by more than seventy percent since June 2014.

The world's environment is affected by the oil price falling. With the drop of oil prices, the fuel bills are lowered. As a result, consumers are very likely to use more oil and thus increase the carbon emission. In addition, there is less incentive to develop renewable and clean energy resources. On the other hand, sustained low oil prices could lead to a drop in global oil and gas exploration and exploitation activities.

Fluctuating oil prices also play an important role in the global economy (Husain et al., 2015). The fall in oil prices would result in a modest boost to global economic activity, although the owners of oil sectors suffer income losses. Recent research from the World Bank shows that for every 30% decline of oil prices, the global GDP (Gross Domestic Product) would be increased by 0.5%. At the same time, the drop of oil prices would reduce the cost of living, and hence the inflation rate would fall.

There is no doubt that crude oil price forecasts are very useful to industries, governments as well as individuals. Thus, forecasting crude oil prices has been the subject of research by both academia and industry. Many methods and approaches have been developed for predicting oil prices. However, due to the high volatility of oil prices (Regnier, 2007), it remains one of the most challenging forecasting problems.

* Corresponding author.
E-mail address: leiyalin@cugb.edu.cn (Y. Lei).
Peer-review under responsibility of China University of Geosciences (Beijing).

http://dx.doi.org/10.1016/j.gsf.2016.08.002

In recent years, machine learning techniques have been used in many applications in geosciences (Alavi et al., 2016). Machine learning provides powerful computational tools and algorithms that can learn from and make predictions on data. In this paper, we propose a novel approach for crude oil price prediction based on a new machine learning paradigm called stream learning (Gama et al., 2009; Bifet et al., 2010). The main advantage of our stream learning approach is that the prediction model can capture the changing pattern of oil prices since the model is continuously updated whenever new oil price data are available, with very small constant overhead. We compare our stream learning model with three other popular oil price prediction models for predicting two types of oil prices (the U.S. refiner acquisition cost for crude oil imports and the WTI crude oil spot price). The experiment results show that our stream learning model achieves the highest accuracy in terms of both mean squared prediction error and directional accuracy ratio over a variety of forecast time horizons.

2. Methods

2.1. Literature review

We divide crude oil price forecasting approaches into three categories: (1) heuristic approaches; (2) econometric models; and (3) machine learning techniques.

Heuristic approaches for oil price prediction include professional and survey forecasts, which are mainly based on professional knowledge, judgments, opinion and intuition. Another heuristic approach, the so-called no-change forecast, uses the current price of oil as the best prediction of future oil prices. Despite its simplicity, the no-change forecast appeared to be a good baseline approach for oil price prediction and was better than other heuristic judgmental approaches (Alquist et al., 2013).

Econometric models are the most widely used approaches for oil price prediction, which include autoregressive moving average (ARMA) models and vector autoregressive (VAR) models, with possibly different input variables (Pindyck, 1999; Frey et al., 2009). These econometric models provide more accurate prediction than the no-change model at least at some horizons (Alquist et al., 2013; Baumeister and Kilian, 2015). Recently, a forecast combination approach was proposed by Baumeister and Kilian (2015), which combines 6 different oil price prediction models including both econometric models (such as the VAR model) and the no-change model. It should be noted that most of the econometric models are linear models and are not be able to capture the nonlinearity of oil prices.

Several machine learning techniques were proposed for oil price prediction, such as artificial neural networks (ANN) (Yu et al., 2008; Kulkarni and Haidar, 2009), and support vector machine (SVM) (Xie et al., 2006). These are nonlinear models which may produce more accurate predictions if the oil price data are strongly nonlinear (Behmiri and Pires Manso, 2013). However, these machine learning techniques, like other traditional machine learning techniques, rely on a fixed set of training data to train a machine learning model and then apply the model to a test set. Such an approach works well if the training data and the test data are generated from a stationary process, but may not be effective for non-stationary time series data such as oil price data.

2.2. A new stream learning approach

In this paper, we propose a new stream learning approach for oil price prediction. Unlike traditional machine learning algorithms that use "one-shot" data analysis and focus on homogeneous and stationary data, stream learning algorithms have been developed to handle applications where continuous data streams are generated from non-stationary processes (Gama et al., 2009; Bifet et al., 2010).

Under the traditional supervised machine learning framework, one typically splits labeled data examples into a training set and a test set. The training set is used to train a machine learning model using a machine learning technique such as ANN or SVM. The test set then is used to test the performance of the machine learning model (note that sometimes there is also a development set which is used to tune the parameters of the machine learning model). For such an approach to be useful, there is an underlying assumption that the data examples in the training set and in the test set are homogeneous (e.g., generated from a stationary process), so that the trained machine learning model can capture the pattern of the data examples in the test set and produce accurate predictions for them. However, oil price time series data are not stationary, and thus traditional machine learning approaches may not produce accurate predictions.

In recent years, a new machine learning paradigm called stream learning has emerged to handle real-world applications where (1) there is a continuous flow of data as opposed to a fixed sample of independent and identically distributed (i.i.d.) examples, and (2) the data are generated by a non-stationary process instead of a stationary process (potentially at very high speed). Examples of stream learning applications include social networks, web mining, scientific data, financial data, etc.

Suppose we start with a set of initial training data which include a sequence of historical oil prices and the associated input data vectors, denoted by $\mathrm{D} = \{(x_{-m+1}, y_{-m+1}), (x_{-m+2}, y_{-m+2}), \ldots, (x_{-1}, y_{-1}), (x_0, y_0)\}$, where y_t is the oil price at time slot t, and x_t is a vector of input data variables for predicting y_t. In this paper, we use the oil prices of the previous o time slots to predict the oil price of time slot t, i.e., $x_t = (y_{t-o}, \ldots, y_{t-2}, y_{t-1})$, where o ranges from 1 to 10. Our framework is applicable to general time slot unit (also called forecast time horizon), which could be daily, weekly, monthly, quarterly, etc.

Suppose we want to predict the oil prices for the next n time slots, denoted by $y_1, y_2, \ldots, y_n$ and again let x_t be input data vector for predicting y_t, $t = 1, \ldots, n$. We propose the following stream learning procedure for predicting oil prices.

Stream learning procedure

1. Use the data in the initial training set D to train a machine learning model, denoted by M_1, and use M_1 to predict the oil price for time slot 1.
2. For time slot t, $t = 2, \ldots, n$:
   Add (x_{t-1}, y_{t-1}) to the training set D and update the machine learning model, denoted by M_t, and use M_t to predict the oil price for time slot t.

Our stream learning approach is a supervised machine learning method which uses a set of labeled training data to train an initial model. The main features and advantages of using stream learning for oil price prediction include:

(1) The machine learning model will be updated whenever new oil price data are available, so the model continuously evolves over time, and can capture the changing pattern of oil prices.
(2) For non-stationary time series data such as oil prices, a forgetting mechanism (e.g., sliding windows, fading factors) will be deployed when updating the machine learning model.
(3) Updating the model requires only a small constant time per new data example, as opposed to re-training the model using the entire training data set.

Please cite this article in press as: Gao, S., Lei, Y., A new approach for crude oil price prediction based on stream learning, Geoscience Frontiers (2016), http://dx.doi.org/10.1016/j.gsf.2016.08.002