系统工程理论与实践
Systems Engineering-Theory & Practice
ISSN 1000-6788,CN 11-2267/N

《系统工程理论与实践》网络首发论文

题目：融合新闻影响力衰减的国际原油价格预测研究
作者：刘岭，王聚杰，李建平
网络首发日期：2022-01-14
引用格式：刘岭，王聚杰，李建平．融合新闻影响力衰减的国际原油价格预测研究[J/OL]．系统工程理论与实践.
https://kns.cnki.net/kcms/detail/11.2267.N.20220113.1807.018.html

doi: 10.12011/SETP2021-2010

融合新闻影响力衰减的国际原油价格预测研究

刘岭[1]，王聚杰[1]，李建平[2]

(1.南京信息工程大学 管理工程学院，南京 210044; 2.中国科学院大学 经济与管理学院，北京 100190)

摘 要 随着大数据技术的发展，新闻数据被应用于国际原油价格的预测，但是目前缺乏对新闻数据影响程度和影响时长的研究. 为量化新闻影响衰减，本文基于指数衰减和互信息提出了一种新闻影响力指数衰减时间序列的计算和择优方法. 为提高预测精度，本文构建了基于差分进化优化算法的组合核函数支持向量回归预测模型，实现了权重系数、核函数参数和回归模型参数的优化选取. 为评估方法的有效性，本文选择了 8 个模型进行对比研究. 实证结果表明：本文设计的新闻影响力指数衰减时间序列计算方法提高了新闻指数与原油价格的相关性，有利于提升原油价格的预测精度；本文设计的预测模型具有较好的预测精度，验证集的平均绝对百分比误差为 1.53%，优于对比模型.

关键词 原油价格；指数衰减；互信息；支持向量回归

International crude oil price forecasting with news influence index attenuation

Liu Ling[1], Wang Jujie[1], Li Jianping[2]

(1. School of Management Science and Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China; 2. School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract With the development of big data technology, news data has been applied to forecast international crude oil price, but there is a lack of research on the impact degree and duration of news data. In order to quantify the decay of news influence, this paper puts forward a new exponential decay analysis method based on exponential decay and mutual information. In order to improve the prediction accuracy, this paper has constructed a combined kernel function support vector regression prediction model based on differential evolution optimization algorithm which realize the optimal selection of weight coefficient, kernel function parameters and regression model parameters. To evaluate the effectiveness of the proposed methods, eight models are selected for comparative study. The empirical results show that integration of news influence decay time series can improve the correlation and prediction accuracy of crude oil price; the proposed forecasting method has good prediction accuracy, the average absolute percentage error of verification set is 1.53%, which is better than the comparison models.

Key Words Crude oil price; Exponential attenuation; Mutual information; Support vector regression

---

作者简介:刘岭(1983-)，男，汉族，河北邯郸人，博士研究生，研究方向：大数据管理决策，E-mail：liulingcn@139.com；通讯作者：王聚杰(1983-)，男，汉族，河北邯郸人，副教授，硕士生导师，博士，研究方向：金融科技与金融风险管理等，E-mail：jujiewang@126.com；李建平(1976-)，男，汉族，浙江建德人，教授，博士生导师，博士，研究方向：风险管理、大数据管理决策，E-mail：ljp@ucas.ac.cn.

基金项目:国家自然科学基金资助项目(71971122，71501101)；江苏省研究生科研与实践创新计划资助项目(KYCX20-0940)

Foundation item: National Natural Science Foundation of China (71971122, 71501101); Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX20-0940)

1 引言

能源金融市场的全球化对世界经济发展、能源利用和环境改善产生了重要影响[1]. 原油是全球能源金融市场的重要大宗商品, 具有商品、金融、战略等多重属性[2]. 原油价格波动对宏观经济具有很大影响, 分析和预测原油价格有助于防范金融风险, 具有重要的理论意义和实际价值[3]. 原油价格受到国际政治、经济危机、军事战争、自然灾害、新冠疫情等多方面因素的影响, 存在波动性、复杂性和多变性[4]. 原油价格的波动也对汇率市场、股票市场、期货市场和全球经济产生影响, 即具有溢出效应[5]. 影响因素导致原油价格的波动, 原油价格的波动产生溢出效应, 溢出效应对影响因素又产生影响. 这种复杂的相互关系导致精确预测原油价格十分困难, 因此众多学者致力于原油价格的预测研究.

随着大数据技术的发展, 与原油价格相关的海量新闻信息被获取、分析、量化为时间序列, 然后融合到原油价格数据[6]. 新闻信息具有实时性和影响力, 其引入丰富了数据源, 有助于提升预测精度. Lucey 等的研究表明新闻可以影响投资者情绪, 且有助于预测原油价格[7]. 新闻数据的量化分析主要采用词频统计和隐含狄利克雷分布(LDA) 方法. 词频统计方法对关键词进行统计, 方法简单, 可解释性强. 张跃军等通过 Google 网站关键词的检索频率构建投资者关注度指数, 应用于原油价格的波动研究[8]. LDA 方法通过对文本信息进行量化分析实现主题和关键词的识别、分类和统计. Li 等利用 LDA 对新闻标题进行分类, 应用于原油价格的预测研究[9]. 现有大多数研究直接利用量化后的新闻时间序列, 建立其与预测对象的点对点映射关系, 以构建出融合新闻指数的数据集. 王晓丹等构建了新闻与股票交易的一一映射关系, 利用两个交易日之间的新闻数据对应当前交易日的股票数据[10].

原油价格的预测研究多采用美国 WTI 原油价格或欧洲 Brent 原油价格数据, 相关预测方法可以分为统计学方法和机器学习方法. 统计学方法具有严格的数学基础, 如最小二乘法回归分析、差分自回归移动平均模型(ARIMA)、向量自回归模型(VAR) [9]、广义自回归条件异方差模型(GARCH) [10] 等, 但是统计学方法在处理大规模、非线性、复杂关系数据时存在局限性 [11]. 机器学习方法具有处理复杂非线性数据的优势, 常用的机器学习方法主要包含支持向量回归(SVR) [12]、反向传播神经网络(BP)和长短记忆神经网络(LSTM) [13]. 支持向量回归(SVR)对非线性小样本时间序列具有较好的预测能力而被广泛应用. 张新生利用 SVR 预测长输油管道的腐蚀深度, 其数据集为 65 个中缅长石油管道测试点的小样本数据[14]. 梁小珍利用 SVR 预测航空客运需求, 其数据为 144 个上海浦东和虹桥两个机场的月度数据[15]. 支持向量回归还具有结构简单, 运算快等优点, 因此适用于原油价格的预测分析. Wang 等利用 SVR 构建了原油价格组合预测模型[16]. 王珏等利用三个不同核函数的 SVR 模型分别预测原油价格[17].

现有研究在新闻数据处理方面存在两个问题, 一是缺少对新闻影响力衰减的量化研究, 二是缺少对新闻影响力衰减与原油价格映射关系的研究. 现实中, 新闻产生的影响不会快速消失, 不仅对当日产生影响, 而且会持续一段时间. Ding 等的研究表明金融新闻长期衰减对股票市场产生影响[18]. 戴光全等的研究发现广交会新闻报道的影响力存在时间衰减现象[19]. 因此不考虑新闻影响的衰减和影响的时长, 就难以准确的反映客观现实. Simkin 等的研究结果表明, 文章在网页上的消失速度和人类行为影响是网络文章关注度下降的原因, 并构建指数衰减方程计算网络新闻关注度衰减指数[20]. 鉴于此, 本文构建了新闻影响力的指数衰减方程, 设计了新的基于互信息的指数衰减时间序列计算方法.

现有关于原油价格的预测研究大多采用单核 SVR 进行预测分析[16], 并通过智能优化算法对 SVR 参数进行优化, 以提高预测精度. 智能优化算法通过初始化种群, 设计随机搜索和个体更新方法, 实现优化过程的收敛[21]. 常用的方法有差分进化法、模拟退火算法、粒子群优化算法、灰狼优化算法等 [22]. 梁小珍等利用粒子群算法优化 SVR 预测航空客运需求[15]. 林玲等基于改进的灰狼优化算法优化 SVR 预测网络舆情[23]. 支持向量机核函数的优化选择和组合构建也可以提高预测精度. 王珏等的预测结果显示基于 RBF 核的 SVR 预测误差小于基于 Linear 核和 Sigmoid 核[17]. 陈家乾等的预测结果显示多核 SVR 预测误差小于单核预测误差[24]. 鉴于此, 本文设计了新的基于差分进化算法的组合核函数预测模型. 通过增加权重系数, 本文将 RBF 核函数、Laplacian 核函数和多项式核函数进行线性组合, 构建出新的组合核函数. 每个核函数的权重系数利用差分进化算法进行优化.

为评估本文方法的有效性，本文利用欧洲 Brent 原油价格数据和 oilprice.com 网站的新闻数据对原油价格进行预测，并选取了 8 个模型进行实证分析. 本文章节安排如下：第 2 节介绍新闻影响力时间序列的量化方法；第 3 节介绍预测方法、模型构建和评估方法；第 4 节进行实证分析；第 5 节是结论.

2 新闻影响力量化方法

2.1 新闻数据的词频统计

本文采用的新闻文本数据源自于原油价格的专门网站，所以不需要对新闻内容的主题进行分析. 为量化新闻的影响力，本文采用了简单有效的词频统计方法，即统计新闻内容中 oil 单词的出现次数. 通过将每日所有新闻文本中含有 oil 单词的个数进行累加，得到每日频次，频次的大小反应了未考虑指数衰减时当日新闻影响力的大小. 例如，第i日词频数k_i为未考虑指数衰减的当日新闻影响力.

2.2 指数衰减方程的构建

参照 Simkin 关于网络新闻呈现指数衰减的研究 [20]和物理学的放射性衰变定律 [25]，本文假设新闻影响力随着时间的推移以指数形式衰减，且数值的衰减速度与当前值成正比. 定义新闻影响力指数衰减方程的一般公式如下:

$$N'(t) = -\lambda N(t) \tag{1}$$

求解微分方程，可得:

$$N(t) = N_0 \mathrm{e}^{-\lambda t} \tag{2}$$

其中$N(t)$为衰减函数，即N_0在 t 时刻的新闻影响力衰减值；t 为时间，$t = 0$时$N(0) = N_0$为初始时间上单位新闻影响力产生的影响；λ为衰减指数，其值大于 0.

为求解衰减系数，本文假定新闻影响力衰减为 0.001 时对应衰减时间区间的终点t_{end}，并将N_0设为 1. 如果新闻在第 m 天的影响衰减为 0.001，即$t_{end} = m$，则可以计算出方程(1)的系数$\lambda = -ln\ 0.001/m$. 不同 m 值对应的λ值也不同，相应的指数衰减曲线也不同. 以衰减时长$m = 4$为例，可计算出$\lambda = 1.727$，次日起连续四天的单位衰减值为 [0.178, 0.032, 0.006, 0.001]，如图 1 所示. 从图 1 中可以看出，m 值越小，指数曲线衰减的速度越快.

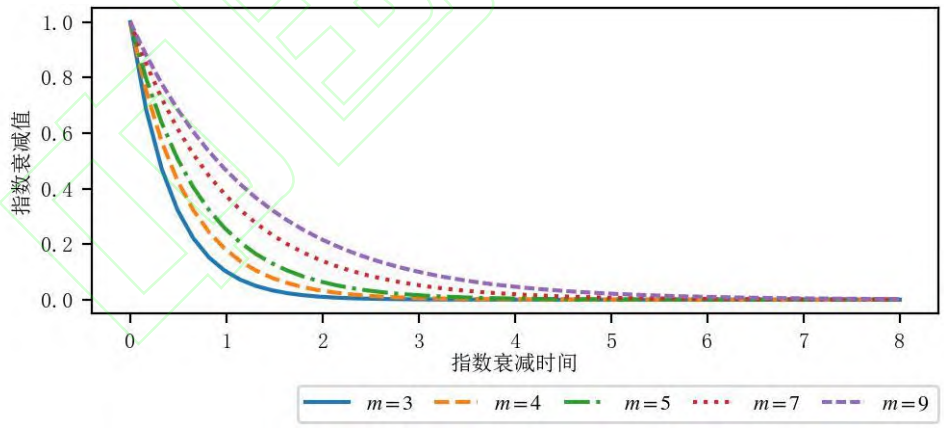

图 1 指数衰减

2.3 衰减时长的选取方法

为选取合适的衰减时长 m，本文设计了基于互信息和指数衰减方程的选取方法. 该方法利用互信息(MI)计算不同衰减时长下，新闻影响力数据与原油价格数据的相关性，并根据互信息的变化趋势，选取最优衰减时长. 互信息计算两个变量之间的联合信息熵，可以反映两个序列之间的非线性相关性大小 [26]，计算公式如下:

$$MI(X,Y) = \iint dxdy\mu(x,y)log\frac{\mu(x,y)}{\mu_x(x)\mu_y(y)} \tag{3}$$

其中$u\ (x,y)$为变量 x 和 y 的联合概率密度,$u_x(x)=\int dy\ \ u(x,y)$和$u_y(x)=\int dx\ \ u(x,y)$为边际概率密度. 互信息为非负值, 当且仅当两个变量相互独立时为 0.

衰减时长m的选取方法和新闻影响力衰减时间序列的计算方法如下:

(1)将单位数据产生的影响设置为 1. 设置 m 的范围, 例如$m\in\ [1,15]$, 即衰减时长为 1 天至 15 天;

(2)求解不同m下λ值, 并计算各个衰减子区间[0, 1], [0, 2]⋯[0, m]的单位衰减时间序列.

(3)针对每个衰减子区间, 每日新闻影响力 k_i 乘以其单位衰减时间序列, 得到的子序列按相同日期求和后得到相应的新闻影响力衰减时间序列.

(4)在相同的m值下, 计算每个子区间的新闻影响力衰减时间序列与原油价格时间序列的互信息, 选取最大互信息值对应的子区间为最优子区间. 最后, 根据最优子区间长度与互信息值的变化, 选择最优的衰减时长m, 其对应的新闻影响力衰减时间序列为所求时间序列.

2.4 新闻影响力与原油价格的映射关系

原油价格交易时间为周一至周五, 因此周六和周日没有数据. 新闻数据也不是每天都有, 可能为零. 数据的映射关系是建立数据集的基础. 本文映射关系的构建方法, 以 2021 年 3 月 15 日至 3 月 22 日的实际数据进行说明, 如图 2 所示. 图 2 中新闻词频数据在 3 月 21 日为 0, 原油价格在 3 月 20 日和 3 月 21 日没有数据. 新闻词频数据经过指数衰减计算后, 得到每日连续的新闻影响力衰减时间序列, 这一映射过程是复杂的多对一过程, 如 2.3 节所述. 新闻影响力衰减时间序列的每日数据, 反映了当日和之前一段时间共同产生的影响. 因此, 将新闻影响力数据按照原油价格数据的日期进行对应, 就可以得到需要的数据集.

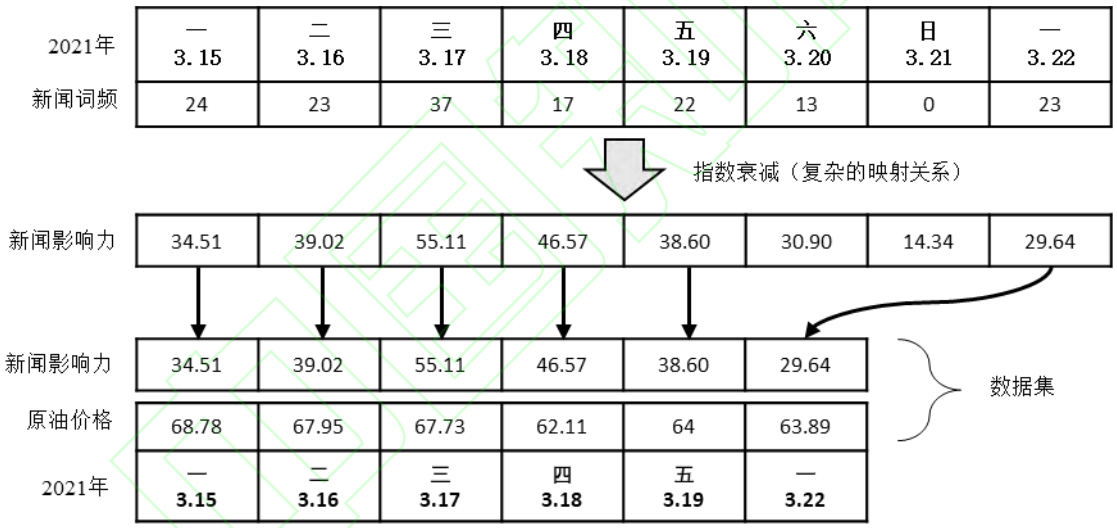

图 2 新闻影响力映射关系

3 预测与评估方法

3.1 组合核函数支持向量机

支持向量机(SVM)源于求解分类问题, 通过求解最大边距超平面对特征子集进行划分. 对于给定训练集$\{(x_1,y_1),\dots,(x_n,y_n)\}\in R^d$, 计算$\varepsilon$偏差下 SVM 的超平面方程为[27]:

$$f(x)=<\omega,x>+b \tag{4}$$

其中$<\ ,>$为点积运算. 超平面的优化方程和约束条件为:

$$\min\quad \frac{1}{2}\|\omega\|^2 \tag{5}$$

$$s.t.$$

$$y_i-<\omega,x_i>-b\le\varepsilon$$

$$<\omega,x_i>+b-y_i\le\varepsilon$$

支持向量回归 SVR 基于 SVM, 计算$f(x)$和y之间的损失, 其优化方程和约束条件为:

$$\min \quad \frac{1}{2}\|\omega\|^2 + \mathrm{C}\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{6}$$

$$s.t.$$
$$y_i - < \omega, x_i > -b \le \varepsilon + \xi_i$$
$$< \omega, x_i > +b - y_i \le \varepsilon + \xi_i^*$$
$$\xi_i, \xi_i^* \ge 0$$

其中, C 为惩罚参数($\mathrm{C} > 0$), ε为不敏感损失系数, ξ_i和ξ_i^*为松驰变量. 为解决非线性问题, 引入核函数将样本集从低维空间映射至高维空间, 即$\mathcal{R} \to \mathcal{F}$, 此时超平面方程为:

$$f(x) = < \omega, \Phi(x) > +b \tag{7}$$

常用的核函数有 RBF 核k_r, 多项式核k_m和 Laplacian 核k_l, 它们的计算公式分别为:

$$k_r(x,y) = \exp(-\gamma \cdot \|x-y\|^2) \tag{8}$$
$$k_m(x,y) = (\gamma \cdot < x, y > +c_0)^d \tag{9}$$
$$k_l(x,y) = \exp(-\gamma \cdot \|x-y\|_1) \tag{10}$$

其中, γ、c_0和 d 为核函数的相关参数, $\|x-y\|_1$为变量之间的 Manhattan 距离. 本文将多项式参数c_0和多项式次数d设为默认值 1 和 3.

SVR 模型的预测精度和泛化能力受 C, ε和γ三个重要参数的影响[28]. C 为惩罚参数($\mathrm{C} > 0$), 反映了算法对超出ε管道样本数据的惩罚程度, C 值变小对超出的样本数据惩罚小, 使训练误差变大, 反之则使模型的泛化能力变差. ε控制着回归函数对样本数据的不敏感区域的宽度, 其值过小使泛化能力变小, 其值过大会导致学习精度不够. γ定义了单个样本的影响范围, 影响数据映射到特征空间的分布.

核函数的设置和选择对 SVR 预测方法至关重要, 组合核函数提供了较好的解决方法. 本文借鉴 Jean-Philippe Vert 提出的组合核函数的设计方法[29], 设计了组合核函数如下:

$$k_c(x,y) = w_1 k_r(x,y) + w_2 k_l(x,y) + w_3 k_m(x,y) \tag{11}$$

其中$k_c(.)$为设计的组合核函数, w_1, w_2, w_3分别为 RBF 核、Laplacian 核和多项式核的权重. 为合理设置 SVR 参数和组合核函数权重, 本文利用差分进化算法对其进行优化.

3.2 差分进化优化算法

差分进化优化算法(Differential evolution optimization algorithm, DE)是一种基于随机种群搜索的全局优化算法[30]. 种群中个体的更新采用随机差分方程, 当更新后的解优于个体当前解时, 个体被更新. 随机差分方程在更新过程中发挥着主要作用, 具有多种形式, 但常用形式为:

$$x_{ni} = x_{r1} + \mathrm{F}(x_{r2} - x_{r3}) \tag{12}$$

其中x_{ni}为第i个种群个体的更新, x_{r1}, x_{r2}, x_{r3}为随机选取的三个种群个体, F 为[0,1)之间的随机数. 优化过程中, $\mathrm{F}(x_{r2} - x_{r3})$产生具有随机方向和随机长度的向量, 提高了全局搜索能力, 加快了收敛速度. 搜索过程中, 算法还通过随机选择更新个体, 修复个体更新过程中超出边界的向量等方法, 增加搜索过程的随机性, 保证个体更新在有效的数值范围.

3.3 数据的标准化

数据在输入预测模型之前, 需要进行标准化处理, 具体公式如下[31].:

$$X_{tr} = \frac{(X - X_{min})}{X_{max} - X_{min}}(S_{max} - S_{min}) + S_{min} \tag{13}$$

其中 X为输入数据, X_{max}和X_{min}分别为其最大和最小值, S_{max}和S_{min}是标准化设置的上限和下限, X_{tr}为标准化后的结果. 预测模型的输出结果需要进行逆标准化转换, 以得到正确的结果, 公式如下:

$$X_{in} = \frac{(X_{tr} - S_{min})(X_{max} - X_{min})}{S_{max} - S_{min}} + X_{min} \tag{14}$$

其中X_{in}为逆标准化转换的结果.

3.4 模型构建

本文基于原油价格数据、新闻文本数据、组合核函数 SVR 和差分进化算法构建预测模型 DE-SVRc, 其结构流程如图 3 所示. 首先, 新闻文本数据经过词频统计和指数衰减处理后, 转化为由数字组成的新闻影响力衰减时间序列, 然后进行数据标准化处理. 原油价格数据经过数据筛选和标准化后, 与对应时间的新闻影响力衰减数据融合成数据集, 用于预测模型的输入.

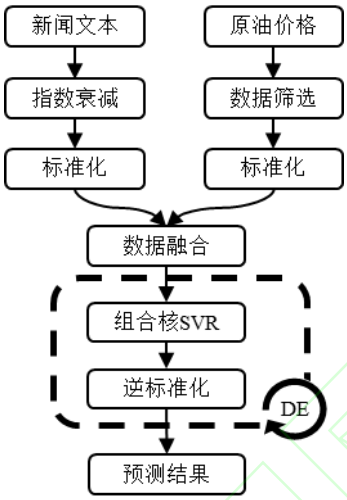

图 3 DE-SVRc 预测模型结构

其次, 将融合的数据集为三部分 [13], 其中训练集占比 70%, 测试集占比 20%, 验证集占比 10%. 训练集和测试集参与预测模型的优化训练, 验证集仅用于验证最终预测结果, 不参与模型的优化训练. 模型采用单步循环预测, 基于前 3 天的数据, 预测第 4 天的数据.

最后, 进行参数优化和结果预测. 组合核 SVR 实现对数据集的训练和预测, 同时利用差分进化优化算法对预测模型参数进行优化. 优化的参数为组合核的权重参数w_1, w_2, w_3, SVR 算法的参数 C, ε和γ. 优化目标为每次预测结果的平均绝对百分比误差$MAPE$, 即通过调整参数使得$MAPE$值最小. 本文采用的是单目标优化, 所以选取了无量纲的$MAPE$指标做为优化目标.

3.5 对比模型

为评估 ED-SVRc 模型, 本文引入 8 个模型 M1-M8 进行对比分析. 模型 M1-M3 分别为基于 RBF 核, Laplacian 核和 3 次多项式核的 SVR. 模型 M5 和 M6 分别采用门控单元神经网络模型(GRU)和长短记忆神经网络模型(LSTM), 都采用 5 层网络结构, 其中 3 个 LSTM 层, 1 个 Dropout 层(减少过拟合)和 1 个 Dense 层(输出预测结果). 模型 M7 采用现有研究 VMD-LSTM-MW 模型 [32], 模型 M8 采用现有研究 EEMD-LSTM 模型 [33]. 模型 M1-M3 用于对比单核 SVR; 模型 M5-M6 用于对比神经网络模型; 模型 M7-M8 用于对比现有研究模型, 没有融合新闻数据. 为对比新闻影响力点对点映射的数据处理方法和本文提出的指数衰减数据处理方法的预测精度, 本文引入对比模型 M4, 此模型除未采用指数衰减处理的新闻数据以外, 其它预测方法与本文提出模型 Mp 相同.

3.6 评价指标

本文采用现有研究中常用的三个评价指标: 平均绝对误差(*MAE*)、均方根误差(*RMSE*)和平均绝对百分比误差(*MAPE*), 相关计算公式如下[34]:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - \hat{x}_i| \tag{15}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|x_i - \hat{x}_i|^2} \tag{16}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - \hat{x}_i}{x_i}\right| \tag{17}$$

其中, x_i为真实值, $\hat{x}_i$为预测值. 评价指标越趋近于 0, 说明预测结果的精度越高.

4 实证分析

4.1 数据收集

原油价格新闻文本数据来源于 www.oilprice.com 网站，本文通过网络爬取方式获得 2011 年 6 月 15 日至 2021 年 4 月 30 日(2552 天)，共 12682 条英文新闻数据. 统计结果显示，新闻数据中 oil 单词的每日词频最大值为 91，最小值为 0.

本文采用的原油数据源于美国能源信息管理局(Energy Information Administration, EIA) 官方网站(www.eia.gov)的欧洲布伦特现货离岸价格(美元/桶)，日期范围与新闻文本数据的日期范围相同. 此时间范围内原油价格的最大值为 128.14，最小值为 9.12，平均值为 73.53，标准差为 27.74.

4.2 数据预处理

新闻文本数据的预测处理主要包含词频统计和指数衰减两个部分. 通过计算不同衰减时长下，其子区间的指数衰减影响时间序列与原油价格时间序列的互信息，选择互信息最大的子区间作为衰减区间的最优值，如图 4 中紫色曲线所示.

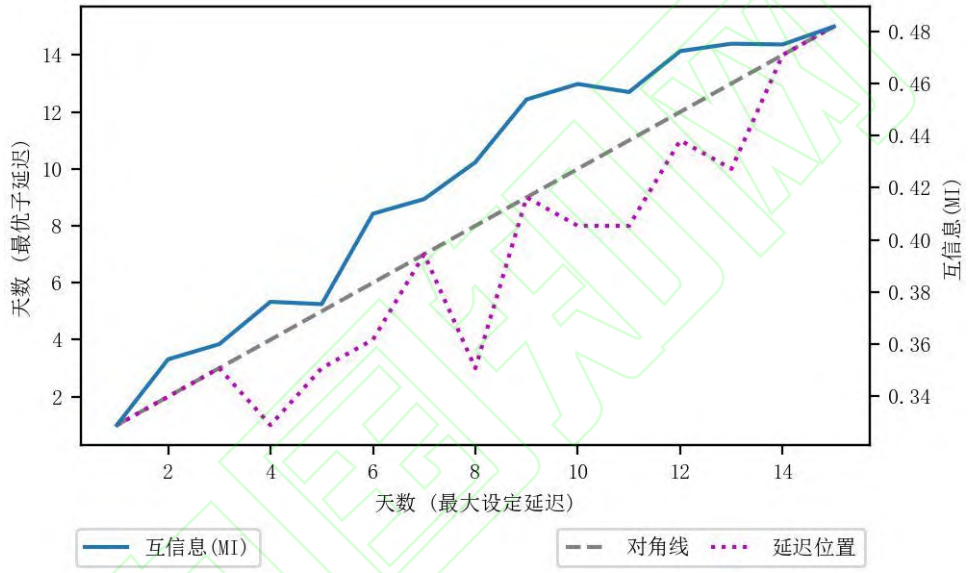

图 4 互信息变化趋势

图 4 中横坐标代表不同衰减时长(最大设定延迟)m，范围从 1 天至 15 天. 以$m = 4$为例，其包含有四个子区间[0,1]，[0,2]，[0,3]，[0,4]，这些子区间内的衰减方程相同，但计算结果不同，分别对应衰减时长为 1 天，2 天，3 天和 4 天的不同衰减时间序列. 这些子区间的衰减时间序列与原油价格序列的互信息计算结果也不同相，互信息最大值对应的区间为[0,1]，对应图 4 中紫色曲线上的(4,1)点. 表明假设$m = 4$时，最优的衰减影响时长不是 4 天，而是 1 天. 图 4 蓝色曲线显示，随着衰减时长的增加，互信息值也增加，但增幅减小，存在波动. 图 4 中最优子区间与m值相同的有 7 个点，分别是$m = 1,2,3,7,9,14,15$，其中$m = 1,2,3,7$ 对应互信息值快速上升，$m = 9$ 之后最优子区间的长度值连续 5 次小于m值，且互信息值增幅放缓. 因此，$m = 9$ 是一个转折点，本文选择其为最优衰减时长.

为说明新闻影响力指数衰减时间序列的变化，本文截取了 2019 年 5 月 1 日至 2021 年 4 月 30 日新闻影响力衰减时间序列(衰减时长为 9 天)和原油价格时间序列的数据，如图 5 所示. 新闻影响力指数衰减时间序列比原始时间序列的幅值增大，且波动变得平缓. 原始的新闻词频时间序列与原油价格时间序列的互信息值为 0.330，增加指数衰减后的互信息值为 0.454，表明指数衰减时间序列与原油价格时间序列之间的相关性较大，显示出指数衰减影响力计算方法的有效性和可行性.

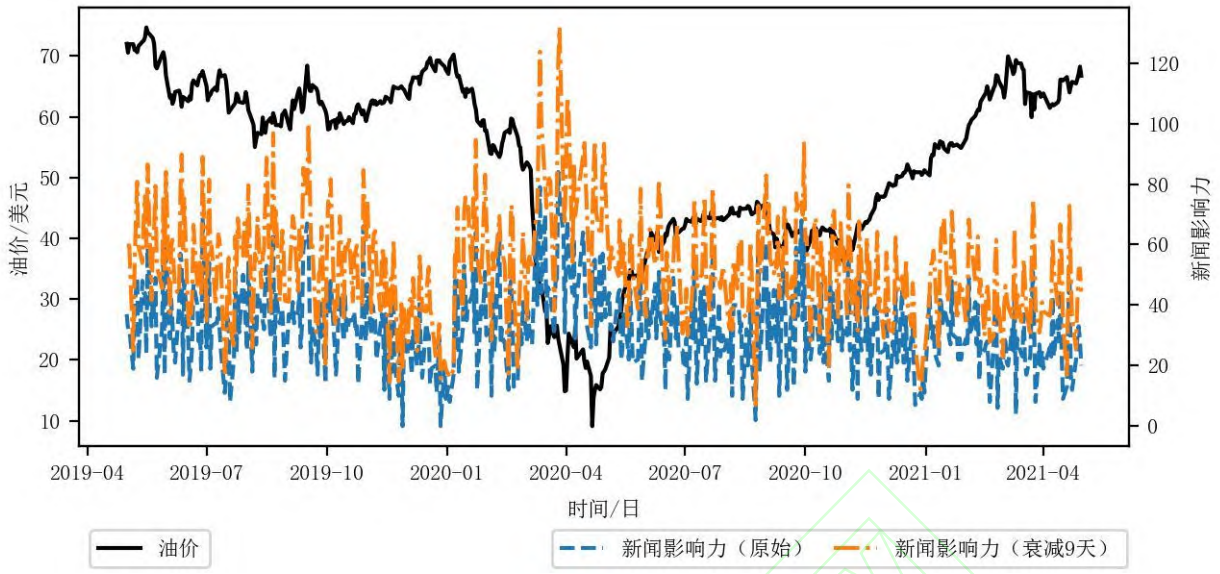

图 5 原油价格、新闻影响力和新闻影响力衰减时间序列

4.3 预测结果与对比分析

根据第 3 节中的方法和模型，本文设计的 ED-SVRc 模型验证集的预测结果如图 6 所示，其中黑色实线为原油价格实际波动曲线，紫色虚线为预测结果. 从图 6 中可以看出，本模型的预测值与实际值的波动变化趋势一致，没有出现明显偏差. 预测结果的误差为$MAE = 0.752$，$RMSE = 1.151$，$MAPE = 1.53\%$，说明本模型的预测精度较高.

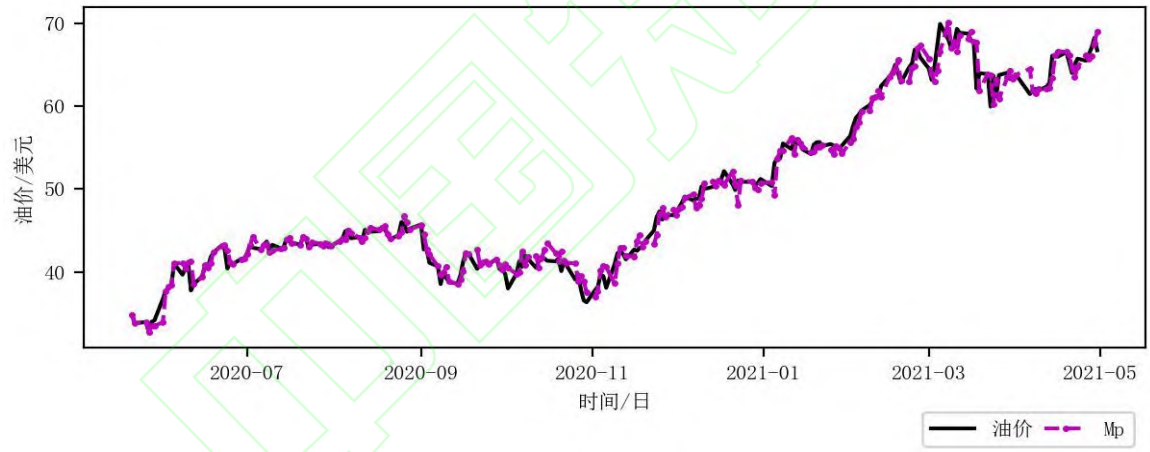

图 6 原油价格和预测结果

为对比评估本文模型的预测结果，本文利用相同的数据对 3.5 小节中的 8 个对比模型进行实验，相关预测结果的误差评价指标如表 1 所示. 为清楚显示各个模型预测结果的曲线，本文选取了验证集的最后 100 个预测数据进行显示，如图 7 所示. 图 7 中将对比模型分两组进行了显示，左边的子图主要显示对比模型 M1 至 M4 的预测结果，右边子图主要显示对比模型 M5 至 M8 的预测结果. 为方便对比，图 7 的两个子图中都包含原油价格曲线和本文模型的预测曲线，前者用黑色实线表示，后者用紫色虚线表示.

分析表 1 中的预测误差数据和图 7 中的预测结果曲线，可以得出四点结论:

第一, SVR 适用于原油价格预测. 第一组对比模型 M1-M4 的预测结果误差较小，其预测曲线的波动趋势和偏差小于第二组对比模型 M5-M8. 这表明基于支持向量机 SVR 的预测方法优于基于神经网络的预测结果，而神经网络预测精度的提升，需要较多的数据对网络进行训练和优化.

第二, "分解-集成"的预测方法可以提高预测精度. 基于"分解-集成"预测方法的 M7 和 M8 模型的预测误差小于直接预测的 M5 和 M6 模型，说明"分解-集成"方法可以提高预测精度. 数据分解以后，子序列与原始序列相比更具趋势性和周期性，这使得子序列的预测难度降低且预测精度提高，经过集成后得到的预

测结果较好.

第三, 组合核函数可以提高 SVR 预测精度. 对比模型 M4 采用了本文设计的模型, 新闻数据没有经过指数衰减处理, 但 M4 的预测误差小于其它对比模型. 本文设计的融合新闻影响力指数衰减的预测模型 ED-SVRc 的预测误差最小. 这两点说明, 组合核函数提高了 SVR 的预测精度.

第四, 新闻数据的指数衰减处理有助于提高预测精度. 增加指数衰减的 Mp 模型的预测精度高于未增加指数衰减的 M4 模型, 表明数据源会影响预测结果. 经过指数衰减处理的新闻影响力数据与原油价格数据的相关性更高, 互信息值较未经过指数衰减的新闻影响力数据高 0.124.

表 1 模型的预测精度对比

模型	名称	*MAE*	*RMSE*	*MAPE*
M_1	SVR_r	0.879	1.194	1.89%
M_2	SVR_l	1.190	1.576	2.57%
M_3	SVR_m	0.857	1.175	1.83%
M_4	ED-SVRc-n	0.849	1.160	1.82%
M_5	GRU	1.664	2.168	3.81%
M_6	LSTM	1.305	1.691	2.89%
M_7	VMD-LSTM-MW	1.147	1.434	2.38%
M_8	EEMD-LSTM	1.419	1.863	2.66%
M_p	ED-SVRc	0.752	1.171	1.53%

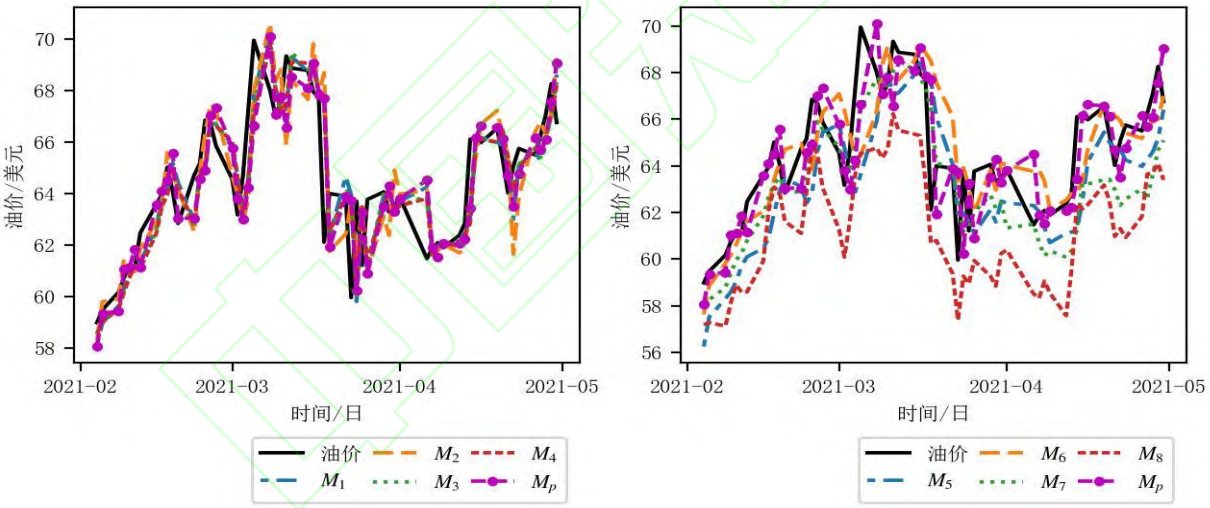

图 7 对比模型的预测结果

5 结论

本文基于指数衰减方程、互信息、差分优化算法和组合核支持向量机等方法, 构建了一种融合新闻影响力衰减的国际原油价格预测模型. 本文针对新闻影响力衰减时间序列计算, 数据融合,预测精度提高等问题, 提出了新的解决方法. 本文选择了8个模型进行对比研究, 实证结果表明: 1) 融合新闻影响力指数衰减时间序列, 有利于提升原油价格的预测精度; 2) 基于组合核函数的SVR预测模型可以提高预测精度; 3) 本文设计的预测模型具有较好的预测精度, 优于对比模型.

国际原油价格的预测研究是有效应对相关金融风险的方法之一. 从国家和政府角度, 预测原油价格可以为相关的财政政策、货币政策、科技政策、能源政策的制定提供依据. 从公司和个人角度, 预测原油价格也可以为原油期货、相关股票产品和理财产品的购买提供参考.

参考文献

[1] 冯钰瑶, 刘畅, 孙晓蕾. 不确定性与原油市场的交互影响测度:基于综合集成的多尺度方法论[J]. 管理评论, 2020, 32(7): 29-40.
Feng Y Y, Liu C, Sun X L. Measuring the interaction between uncertainty and crude oil market: A multiscale methodology based on synthetic integration[J]. Management Review, 2020, 32(7): 29-40.

[2] 刘映琳, 刘永辉, 鞠卓. 国际原油价格波动对中国商品期货的影响——基于多重相关性结构断点的分析[J]. 中国管理科学, 2019, 27(2): 31-40.
Liu Y L, Liu Y H, Ju Z. The impact of international crude oil price fluctuation on Chinese commodity futures --- based on the correlation structure breakpoint model[J]. Chinese Journal of Management Science, 2019, 27(2): 31-40.

[3] 杨坤, 于文华, 魏宇. 基于 R-vine copula 的原油市场极端风险动态测度研究[J]. 中国管理科学, 2017, 25(8): 19-29.
Yang K, Yu W H, Wei Y. Dynamic measurement of extreme risk among various crude oil markets based on R-vine copula[J]. Chinese Journal of Management Science, 2017, 25(8): 19-29.

[4] 黄晓薇, 郭红玉, 郭甦, 等. 国际油价冲击对汇率体系稳定性影响的研究综述与展望[J]. 系统工程理论与实践, 2014, 34(S1): 100-105.
Huang X W, Guo H Y, Guo S, et al. New developments of research on stability of exchange rate under the impact of international oil prices[J]. Systems Engineering — Theory & Practice, 2014, 34(S1): 100-105.

[5] 陈宇峰, 朱志韬, 屈放. 国际油价, 人民币汇率与国内金价的非对称溢出及动态传导机制--基于三元 VAR-Asymmetric BEKK (DCC)--GARCH (1,1)模型[J]. 系统科学与数学, 2021, 41(2): 449-465.
Chen Y F, Zhu Z T, Qu F. Asymmetric spillover effect and dynamic correlation between crude oil, RMB exchange rate and Chinese gold price: based on VAR-Asymmetric BEKK (DCC)-GARCH (1, 1) Model[J]. Journal of Systems Science and Mathematical Sciences, 2021, 41(2): 449-465.

[6] Yu L A, Zhao Y Q, Tang L, et al. Online big data-driven oil consumption forecasting with Google trends[J]. International Journal of Forecasting, 2019, 35(1): 213-223.

[7] Lucey B, Ren B. Does news tone help forecast oil?[J]. Economic Modelling, 2021, 104: 105635.

[8] 张跃军, 李书慧. 投资者关注度对国际原油价格波动的影响研究[J]. 系统工程理论与实践, 2020, 40(10): 2519-2529.
Zhang Y J, Li S H. The impact of investor attention on international crude oil price volatility[J]. Systems Engineering — Theory & Practice, 2020, 40(10): 2519-2529.

[9] Li X R, Shang W, Wang S Y. Text-based crude oil price forecasting: A deep learning approach[J]. International Journal of Forecasting, 2019, 35(4): 1548-1560.

[10] 王晓丹, 尚维, 汪寿阳. 互联网新闻媒体报道对我国股市的影响分析[J]. 系统工程理论与实践, 2019, 39(12): 3038-3047.
Wang X Y, Shang W, Wang S Y. The effects of online news on the Chinese stock market[J]. Systems Engineering — Theory & Practice, 2019, 39(12): 3038-3047.

[11] Klein T, Walther T. Oil price volatility forecast with mixture memory GARCH[J]. Energy Economics, 2016, 58: 46-58.

[12] Valente J M, Maldonado S. SVR-FFS: A novel forward feature selection approach for high-frequency time series forecasting using support vector regression[J]. Expert Systems with Applications, 2020, 160: 113729.

[13] 余乐安, 范常容, 马月明. 基于混合多分形小波的国际油价多期预测研究[J]. 计量经济学报, 2021, 1(3): 612-623.
Yu L A, Fan C Y, Ma Y M. Multi-period prediction of oil price with a hybrid multifractal wavelet model[J]. China Journal of Econometrics, 2021, 1(3): 612-623.

[14] 张新生, 张琪. 基于改进 RFFS 和 GSA-SVR 的长输油管道腐蚀深度预测研究[J]. 系统工程理论与实践, 2021, 41(6): 1598-1610.
Zhang X S, Zhang Q. Research on prediction of corrosion depth of long oil pipelines based on improved RFFS and GSA-SVR[J]. Systems Engineering — Theory & Practice, 2021, 41(6): 1598-1610.

[15] 梁小珍, 郭战坤, 张倩文,等. 基于奇异谱分析的航空客运需求分析与分解集成预测模型[J]. 系统工程理论与实践, 2020, 40(7): 1844-1855.
Liang X Z, Guo Z K, Zhang Q W, et al. An analysis and decomposition ensemble prediction model for air passenger demand based on singular spectrum analysis[J]. Systems Engineering — Theory & Practice, 2020, 40(7): 1844-1855.

[16] Wang J, Zhou H, Hong T, et al. A multi-granularity heterogeneous combination approach to crude oil price forecasting[J]. Energy Economics, 2020, 91:104790.

[17] 王珏, 胡蓝艺, 齐琛. 基于网络关注度的大宗商品市场预测研究[J]. 系统工程理论与实践, 2017, 37(5): 1163-1171.
Wang J, Hu L Y, Qi C. Prediction of bulk commodities based on Internet concerns[J]. Systems Engineering — Theory & Practice, 2017, 37(5): 1163-1171.

[18] Ding X, Shi J H, Duan J W, et al. Quantifying the effects of long-term news on stock markets on the basis of the multikernel Hawkes process[J]. Science China Information Sciences. 2021,64: 192102.

[19] 戴光全, 谭健萍. 基于报纸媒体内容分析和信息熵的广交会综合影响力时空分布[J]. 地理学报, 2012, 67(8): 1109-1124.
Dai G Q, Tan J P. The spatial-temporal distribution of integrated impact index for the canton fair: a study based on content analysis method (CAM) & the information entropy of Chinese Complete Newspaper Database (CCND)[J]. Acta Geographica Sinica, 2012, 67(8): 1109-1124.

[20] Simkin M V, Roychowdhury V P. Why Does Attention to Web Articles Fall with Time?[J]. Journal of the Association for Information Science & Technology, 2015, 66: 1847-1856.

[21] 吴泽忠, 宋菲. 基于改进螺旋更新位置模型的鲸鱼优化算法[J]. 系统工程理论与实践, 2019, 39(11): 2928-2944.
Wu Z Z, Song F. Whale optimization algorithm based on improved spiral update position model[J]. Systems Engineering — Theory & Practice, 2019, 39(11): 2928-2944.

[22] 李彤,王众托.模拟植物生长算法的原理及应用[J].系统工程理论与实践, 2020, 40(5): 1266-1280.
Li T, Wang Z T. The theory and application of plant growth simulation algorithm[J]. Systems Engineering — Theory & Practice, 2020, 40(5): 1266-1280.

[23] 林玲, 陈福集, 谢加良, 等. 基于改进灰狼优化支持向量回归的网络舆情预测[J]. 系统工程理论与实践, 2021, 41(10): 1-11.
Lin L, Chen F J, Xie J L, et al. Prediction of network public opinion based on improved grey wolf optimized support vector machine regression[J]. Systems Engineering — Theory & Practice, 2021, 41(10): 1-11.

[24] 陈家乾, 肖艳炜, 李 英, 等. 多变量相空间重构的多核最小二乘支持向量机电力负荷预测优化策略[J]. 科学技术与工程, 2020, 20(29) : 11956-11962.
Chen J Q, Xiao Y W, Li Y, et al. Multiple kernels LS-SVM power load forecasting optimization strategy based on multiple variate PSR[J]. Science Technology and Engineering, 2020, 20(29) : 11956-11962.

[25] Bakac M, Tasoglu A K, Uyumaz G. Modeling radioactive decay[J]. Procedia - Social and Behavioral Sciences, 2011, 15: 2196--2200.

[26] Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information[J]. Physical Review E, 2004, 69(6): 066138.

[27] Smola A J, Schölkopf B. A tutorial on support vector regression[J]. Statistics and Computing, 2004, 14: 199--222.

[28] Hu G, Xu Z Q, Wang G R, Zeng B, et al. Forecasting energy consumption of long-distance oil products pipeline based on improved fruit fly optimization algorithm and support vector regression[J]. Energy, 2021, 224: 120153

[29] Vert J P, Tsuda K , Schölkopf B, "A Primer on Kernel Methods" in Kernel Methods in Computational Biology[M]. MIT Press, 2004, pp.35-70.

[30] Storn R, Price K. Differential Evolution: A Simple and Efficient Adaptive Scheme for Global Optimization Over Continuous Spaces[J]. Journal of Global Optimization, 1995, 23(1).

[31] Liu L, Wang J J. Super multi-step wind speed forecasting system with training set extension and horizontal–vertical integration neural network[J]. Applied Energy, 2021, 292: 116908.

[32] Huang Y S, Deng Y. A new crude oil price forecasting model based on variational mode decomposition[J]. Knowledge-Based Systems, 2021, 213: 106669.

[33] Wu Y X, Wu Q B, Zhu J Q. Improved EEMD-based crude oil price forecasting using LSTM networks[J]. Physica A: Statistical Mechanics and its Applications, 2019,516: 114--124.

[34] Wu B R, Wang L, Lü S X, et al. Effective crude oil price forecasting using new text-based and big-data-driven model[J]. Measurement, 2021, 168: 108468.