文章编号：1004-4353(2022)04-0347-07

基于多频优化组合模型的我国大豆期货价格预测

郭倩倩， 王星惠， 张从巧
（安徽大学 经济学院，合肥 230601）

摘要：针对 ARIMA、BPNN、LSTM 等单一模型在预测大豆期货价格时因不能同时捕获到原始序列中线性和非线性变化特征而导致的预测精度不高的问题，提出基于完全自适应噪声集合经验模态分解(CEEMDAN)的多频优化组合模型，并利用大豆的日期货收盘价数据对多频优化组合模型的有效性进行了实证分析. 结果表明，多频优化组合模型在大豆期货价格预测精度上优于 BPNN、LSTM 等单一模型，以及 EMD-BPNN、CEEMDAN-LSTM(未重构)等组合模型，因此该模型在预测大豆期货价格走势中具有良好的参考价值.

关键词：大豆期货；价格预测；CEEMDAN 分解；多频优化组合模型
中图分类号：F720 **文献标识码**：A

Prediction of soybean futures price in China based on multi-frequency optimal combination model

GUO Qianqian， WANG Xinghui， ZHANG Congqiao
(*School of Economics*, *Anhui University*, *Hefei* 230601, *China*)

Abstract: To address the problem that single models such as ARIMA, BPNN, and LSTM cannot capture both linear and nonlinear variation in the original series of soybean futures prices, a multi-frequency optimal combination model based on complete ensemble empirical modal decomposition with adaptive noise (CEEMDAN) is proposed. The empirical results show that the multi-frequency optimal combination model outperforms single models such as BPNN and LSTM, as well as combined models such as EMD-BPNN and CEEMDAN-LSTM (unreconstructed) in predicting soybean futures price trends. Therefore, the model has good reference value in forecasting soybean futures price movements.

Keywords: soybean futures; price forecast; CEEMDAN decomposition; multi-frequency optimal combination model

0 引言

大豆不仅是我国重要的粮食作物之一，也是我国进口量最大的农产品，因此大豆在国民经济中占有重要地位. 自 1993 年我国建立大豆期货市场以来，交易量已从初始的 860.69 万手增加到 72.69 亿手(截至 2021 年底)，为稳定我国大豆价格和粮食安全提供了重要保障. 近年来，受国内外多种因素的影响，大豆期货价格处于总体上涨的波动趋势中，因此建立有效的大豆期货价格预测

收稿日期：2022-05-12
基金项目：中国博士后科学基金面上资助项目(2019M662146)；安徽省哲学社会科学规划项目(AHSKQ2020D63)
第一作者：郭倩倩(1996—)，女，硕士研究生，研究方向为金融统计.
通信作者：王星惠(1985—)，男，博士，副教授，研究方向为稳健统计理论方法与应用.

模型对规避大豆价格波动风险和保证我国粮食安全具有重要意义.目前,大豆价格预测方法主要分为两类:一种是利用时间序列模型的线性预测方法(ARIMA[1]、VAR[2]、GARCH[3]等模型)进行预测.该类方法具有操作和计算简单的优点,但对数据的要求较高,而且对非线性时间序列预测的精度和稳定性较低.另一种是利用机器学习的方法(SVR[4]、ANN[5]、LSTM[6]等模型)进行预测.该类方法虽然操作相对复杂,但由于其具有强大的非线性趋势拟合能力(可以更好地捕获数据之间的潜在关联),因此可有效挖掘数据.

近年来,分解集成方法(如EMD、EEMD、CEEMDAN等)因能够更好地捕获到原始序列中不同粒度的特征信息而受到学者的广泛关注,并被应用于多种产品的价格预测中[7-14].2020年,贺毅岳等[15]针对股票市场指数提出了一种基于CEEMDAN-LSTM的预测模型.该模型首先将原始序列分解为若干特征明显的简单序列,然后再利用单一LSTM模型捕获序列在不同频率下的波动性特征.研究显示,该方法可大大改善构建模型的效率和预测精度.但由于不同频率序列在实际中常会表现出不同的趋势,因此单一模型往往难以同时捕获到序列中的其他线性或者非线性特征.为此,本文构建一种基于CEEMDAN的多频优化组合模型,即通过不同单一模型的特性以充分提取不同频率序列的波动特征,以此进一步提高模型的预测精度.

1 模型理论与流程

1.1 CEEMDAN分解方法

CEEMDAN是一种基于EEMD改进的分解方法,它可以有效解决EEMD分解中因添加白噪声序列而导致的计算复杂度增加的问题.

CEEMDAN分解方法的具体步骤为:

1)首先确定第i次添加的白噪声$\omega_i(t)$和幅值ε_k,然后在原始序列中加入白噪声$\omega_i(t)\varepsilon_k$后对其进行EMD分解,并将分解得到的多个子序列IMF均值作为第1阶段的$c\mathrm{IMF}_1(t)$,即:

$$c_1\mathrm{IMF}_1(t)=\frac{1}{I}\sum_{i=1}^{I}E_1[x(t)+\omega_i(t)\varepsilon_0]. \tag{1}$$

2)计算第1阶段残差$r_1(t)$,其计算公式为:

$$r_1(t)=x(t)-c\mathrm{IMF}_1(t). \tag{2}$$

3)将经过EMD分解后的噪声和第1阶残差相加,得到一个新的序列.对该序列进行EMD分解并求集合平均值即可得第2阶段的$c\mathrm{IMF}_2(t)$:

$$c\mathrm{IMF}_2(t)=\frac{1}{I}\sum_{i=1}^{I}E_1[x(t)+E_1(\omega_i(t)\varepsilon_1)]. \tag{3}$$

4)计算第k阶残差$r_k(t)$,其计算公式为:

$$r_k(t)=r_{k-1}(t)-c\mathrm{IMF}_k(t). \tag{4}$$

5)重复第3步,由此可得到第$k+1$阶段的$c\mathrm{IMF}_{k+1}(t)$,即:

$$c\mathrm{IMF}_{k+1}(t)=\frac{1}{t}\sum_{i=1}^{l}E_1[r_k(t)+E_k(\omega_i(t))\varepsilon_k]. \tag{5}$$

6)重复第4步,且当残差序列不可再分解时,将其记为最终的残差($r(t)$),即:

$$r(t)=x(t)-\sum_{k=1}^{K}\mathrm{IMF}_k(t). \tag{6}$$

原始序列$x(t)$的最终分解结果可表示为:

$$x(t)=\sum_{k=1}^{K}\mathrm{IMF}_k(t)+r(t). \tag{7}$$

1.2 预测方法

1.2.1 ARIMA(p,d,q)模型

ARIMA(p,d,q)模型是将自回归模型(AR)、移动平均模型(MA)和差分法相融合而成的一种模型,其中p为自回归向,d是数据进行差分的阶数,q为移动平均项数.ARIMA(p,d,q)模型的表达式为:

$$y_t=\theta_1y_{t-1}+\cdots+\theta_py_{t-p}+\varepsilon_t-\theta_1\varepsilon_{t-1}-\cdots-\theta_p\varepsilon_{t-p}. \tag{8}$$

1.2.2 SVR模型

SVR模型是一种假设能容忍预测值和标签值之间偏差最多为ε的回归模型.如图1所示,该模型通过在线性函数$f(x)$两侧构建一个偏差为ε的隔离带,以最小化偏差ε与总损失来最优化模型.该模型可表示为:

$$\min_{w,b}\ \frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}l_\varepsilon(f(x_i)-y_i),$$

$$l_\varepsilon(z)=\begin{cases}0, & \text{if } |z|<\varepsilon;\\ |z|-\varepsilon, & \text{if } |z|\geqslant\varepsilon.\end{cases} \tag{9}$$

其中:C为正则化常数;l_ε是ε-不敏感损失函数,即损失函数的计算仅针对隔离带外的样本.

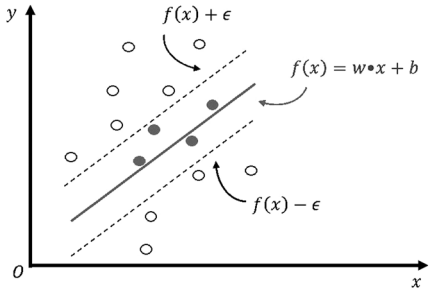

图 1　SVR 的原理示意图

1.2.3　BPNN 模型

BPNN 是一种通过误差反向传播算法训练的多层前馈神经网络模型，其网络结构由输入层、隐藏层、输出层组成，如图 2 所示.

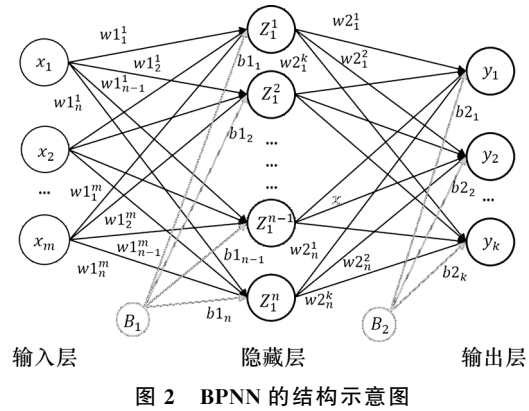

图 2　BPNN 的结构示意图

BPNN 的训练过程如图 3 所示. 训练时：首先根据定义好的损失函数计算预测值和真实值之间的误差，并对其进行求导；然后沿着梯度最小的方向反向传播误差，以此更新网络中每一层的权重参数；最后再进行正向计算. 循环反复此过程，直到损失函数值趋于稳定.

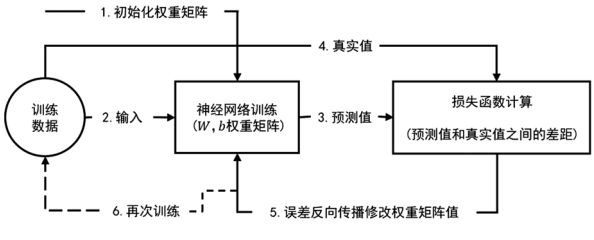

图 3　BPNN 的训练过程

1.2.4　LSTM 模型

LSTM 是一种特殊循环神经网络模型，它可有效解决简单循环神经网络中存在的梯度爆炸或消失的问题. LSTM 网络结构由单元状态、遗忘门、记忆门和输出门组成，如图 4 所示.

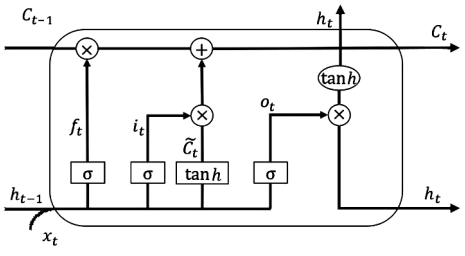

图 4　LSTM 网络结构图

LSTM 的前向计算过程如下：

1) 将当前的输入 x_t 和上一个状态传递下来的 h_{t-1} 进行拼接，并以此作为输入.

2)对长期状态进行控制，主要分为 3 个阶段：

第 1 阶段为忘记阶段. 该阶段主要对上一个节点传来的输入进行选择性忘记，并将输出信号 f_t 作为忘记门控. f_t 的计算公式为：

$$f_t=\sigma(W_f\cdot(h_{t-1},x_t)+b_f), \tag{10}$$

其中 σ 为 sigmod 神经网络层，f_t 为 0 到 1 之间的数(1 代表信息完全保留，0 代表信息完全遗忘).

第 2 阶段为选择记忆阶段. 该阶段主要对输入进行选择性记忆，并由 sigmod 神经网络层的输出 i_t(决定哪些信息需要记忆，由公式(11) 计算所得) 和 tanh 神经网络层的输出 $\widetilde{C}_t$(为整合输入的 h_{t-1} 和 x_t，由公式(12) 计算所得) 共同决定记忆门的输出.

$$i_t=\sigma(W_i\cdot(h_{t-1},x_i)+b_i). \tag{11}$$

$$\widetilde{C}_t=\tanh(W_c\cdot(h_{t-1},x_t)+b_c). \tag{12}$$

第 3 阶段为更新阶段. 模型得到遗忘门和记忆门后更新单元状态(根据公式(13))，t 时刻的单元状态 C_t 可表示为：

$$C_t=f_t\odot C_{t-1}+i_t\odot\widetilde{C}_t, \tag{13}$$

式中 $\odot$ 表示哈达玛积.

3) 根据公式(14) 计算 O_t，并将得到的 O_t 作为输出门控. 根据公式(15) 计算 h_t，并将得到的 h_t 作为下一时刻的输入信号并传递到下一时刻.

$$O_t=\sigma(W_o(h_{t-1},x_t)+b_o), \tag{14}$$

$$h_t=O_t\odot\tanh(C_t). \tag{15}$$

1.3　预测流程

构建对大豆期货收盘价进行预测的多频优化组合模型的方法为：首先，对大豆收盘价进行 CEEMDAN 分解，由此得到 6 个不同频率的 IMF 分量和 1 个残差趋势项；其次，求出各个 IMF 分

量之间的皮尔逊相关系数,并基于皮尔逊相关系数对各个IMF分量进行聚类;再次,对聚类得到的高-中-低频分项和趋势项进行经济意义的解释,并运用SVR、ARIMA、LSTM、BPNN模型对各个分项进行预测,以此选择出各个分项中预测效果最好的模型;最后,组合各个分项的结果,由此得到最终的预测结果.组合模型的预测流程如图5所示.

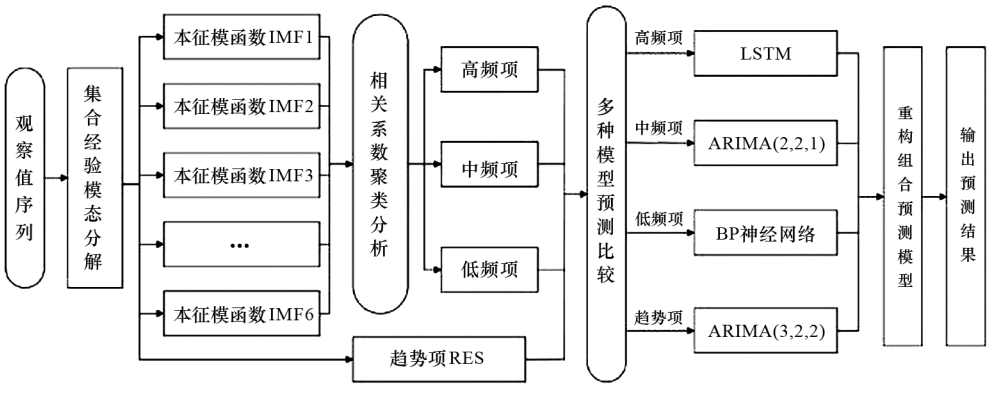

图5 组合模型的预测流程

1.4 模型的评价指标

评价指标采用均方根误差(RMSE)、平均绝对误差(MAE)和对称平均绝对百分比误差(SMAPE),各评价指标的值越小,表明模型的精度越高.各评价指标的计算公式为:

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i-Y_i')^2}, \tag{16}$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|Y_i-Y_i'|, \tag{17}$$

$$SMAPE=\frac{100\%}{n}\sum_{i=1}^{n}\frac{|Y_i'-Y_i|}{(|Y_i'|+|Y_i|)/2}. \tag{18}$$

其中:Y_i表示实际值,Y_i'表示预测值,n表示预测期数.

2 实证分析

2.1 数据来源与选取

本文数据来源于新浪财经网(https://finance.sina.com.cn/),样本区间为2009年8月27日至2021年7月6日的所有交易数据.在数据中剔除日成交量为0的所有收盘价格之后,最后收集到的数据为2 880条.为了减少因数据量较大而产生噪音,本文采用重采样技术(降采样)对数据进行预处理.经预处理后共得到614条数据.图6为数据处理后的大豆期货价格的序列走势图.实验时,本文选取数据集的前80%数据作为训练集,后20%数据作为测试集.

图6 大豆期货价格的序列走势

2.2 大豆期货收盘价的分解和重构

本文运用Python3.7软件对大豆期货日价格进行了CEEMDAN分解,由此共分解得到了6个不同频率的本征模函数(IMF1-IMF6)和1个残差趋势项,如图7所示.

利用Numpy库中的corrcoef函数计算出的IMF1至IMF6之间的皮尔逊相关系数见表1.基于皮尔逊相关系数对各个IMF分量进行聚类的结果见图8.

由表1可以看出,IMF1和IMF3、IMF2和IMF5、IMF2和IMF6、IMF4和IMF6之间的相关系数在1%水平下显著.由图8可以看出:IMF1、IMF2、IMF3和IMF4为高频项,IMF5为中频项,IMF6为低频项.

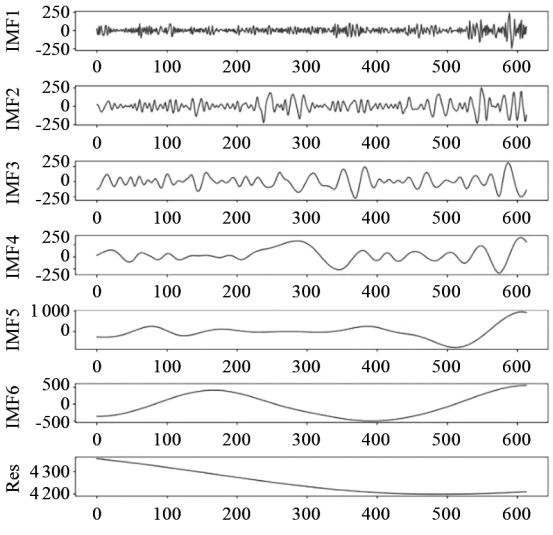

图 7　大豆价格的 CEEMDAN 分解结果

表 1　皮尔逊相关系数值

	IMF1	IMF2	IMF3	IMF4	IMF5	IMF6
IMF1	1					
IMF2	−0.015	1				
IMF3	0.007	0.089	1			
IMF4	−0.014	−0.051	−0.022	1		
IMF5	0.065	−0.001	−0.020	−0.066	1	
IMF6	0.062	0.002	0.087	0.002	0.250	1

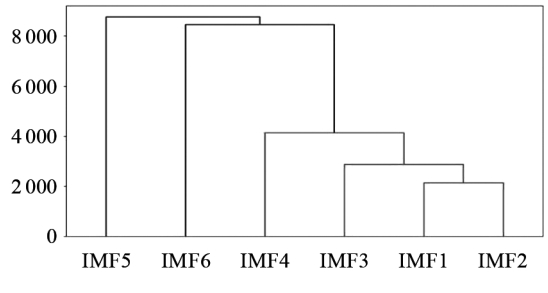

图 8　基于皮尔逊相关系数的聚类图

为分析各分项的特征，利用 Python 科学计算包计算了各个分项的 p 值、周期、方差贡献率以及各分项与原序列的相关系数，结果见表 2. 由表 2 可知：

1）高频项（IMF1-IMF4）的平均周期为 4.451，其与原序列的相关系数为 0.284，p 值为 0.000（远低于 0.05 的水平，即通过显著性检验），表明原序列和高频项之间存在相关性，但相关程度较弱；高频项的方差贡献率为 9.307%，表明高频项对大豆期货收盘价的解释力较小. 图 9 为大豆期货收盘价和各分项重构后的走势. 由图 9 可以看出，高频项的均值始终在 0 附近上下波动，这是由市场短期不规则事件引发的价格变化导致的.

表 2　各分项的周期和方差贡献率

分项名称	与原序列相关系数	p 值	周期	方差贡献率/%
高频	0.284	0.000	4.451	9.307
中频	0.783	0.000	169.600	53.216
低频	0.727	0.000	636.000	36.150
趋势项	0.188	0.011	—	1.328

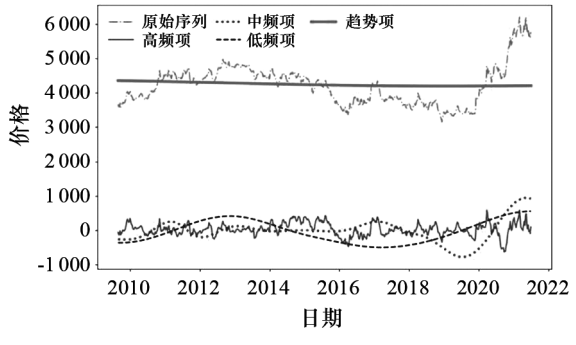

图 9　大豆期货收盘价和各分项重构后的走势

2）中频项（IMF5）的平均周期为 169.600，其与原序列的相关系数为 0.783，p 值为 0.000（通过显著性检验），表明原序列和中频项之间存在相关性，且相关程度较强；中频项的方差贡献率为 53.216%，表明中频项对大豆期货收盘价的解释力较强. 从图 9 可以看出，中频项与原序列的波动基本保持一致，该结果同时也验证了上述分析结论. 此外，从图中还可以看出中频项能够反映出因重大事件导致的大豆价格变化，如 2016 年 12 月国务院推出农业补贴政策转型后，大豆期货价格总体呈现下降趋势（2017 年 1 月 10 日至 2019 年 12 月 3 日）.

3）低频项（IMF6）的平均周期为 636.000，其与原序列的相关系数为 0.727，p 值为 0.000（通过显著性检验），表明原序列和低频项之间存在相关性，且相关程度较强；低频项的方差贡献率为 36.150%，表明低频项对大豆期货价格也具有一定的解释力.

4）残差趋势项不存在周期性，其与原序列的相关系数为 0.188，p 值为 0.011（通过显著性检验），表明原序列和趋势项之间存在相关性，但相关程度弱. 残差趋势项的方差贡献率为 1.328%，表明趋势项对大豆收盘价的解释力较小. 由图 9

可知,大豆期货价格虽然存在较大波动,但会逐渐回到趋势价格附近.这表明,趋势项可以反映大豆期货价格的长期趋势.

2.3 大豆期货收盘价的预测和对比

用SVR、BPNN、ARIMA、LSTM 4个模型对分解重构后的各个分项进行预测,结果见表3.从表3可以看出,在高频项中评价LSTM模型的3个指标几乎都取得了最优的结果.其原因是LSTM模型在面对波动性较强的高频序列时可以考虑到之前序列的特征信息,进而使得模型可以有效捕获到序列中的非线性特征.因此,本文选择LSTM模型作为高频项的预测模型.

表3 4种模型对各分项的预测结果

模型	指标	高频	中频	低频	趋势项
SVR	RMSE	108.110	318.058	113.707	16.321
	MAE	78.823	263.172	108.536	16.276
	SMAPE	76.076	61.347	60.880	0.385
BPNN	RMSE	101.593	7.784	0.121	0.442
	MAE	74.875	6.801	0.104	0.436
	SMAPE	72.165	2.137	0.540	0.010
ARIMA	RMSE	96.693	0.225	3.185	0.000
	MAE	70.216	0.193	2.809	0.000
	SMAPE	146.424	128.858	80.870	0.125
LSTM	RMSE	98.048	27.649	33.793	0.832
	MAE	69.977	18.198	31.146	0.699
	SMAPE	67.319	4.353	16.806	0.016

在中频项的预测结果中,ARIMA模型虽然在SMAPE评价指标上低于其他3个模型,但在RMSE和MAE评价指标上显著优于其他3个模型;因此,经综合考虑后,本文选择ARIMA模型作为中频项的预测模型.

在低频项的预测结果中,BPNN模型的各项评价指标均优于其他模型,因此本文选择BPNN模型作为低频项的预测模型.

对比表3中各单一模型在趋势项上的各评价指标结果可知,除了SVR模型外其他模型均取得较好的预测效果.由于ARIMA模型在捕获趋势项的线性变化上具有更明显的优势,因此本文选择ARIMA模型作为趋势项的预测模型.

基于上述分析,本文集成LSTM、ARIMA(2,2,1),BPNN、ARIMA(3,2,2)模型构建了多频优化模型.图10为模型在测试集上的预测效果图.由图10可看出,多频优化组合模型能够很好地捕获到大豆期货收盘价的波动规律,即模型的拟合能力较强,由此表明模型具有良好的预测效果.

为验证本文构建的多频优化组合模型的优越性,设计了多组对比实验.实验结果如表4所示.由表4可知,除EMD-LSTM组合模型外,其他基于CEEMDAN分解的组合模型的预测效果均显著优于各单一模型.其原因是CEEMDAN能够克服EMD在分解过程中出现的自适应性差及模态混叠的问题.此外,在基于CEEMDAN分解的组合模型中,多频优化组合模型(重构)的预测效果显著优于CEEMDAN-BPNN和CEEMDAN-LSTM未重构组合模型,这表明经分解重构后的模型在预测精度上更具有优势:因此,本文提出的多频优化组合模型更适用于大豆期货价格的预测.

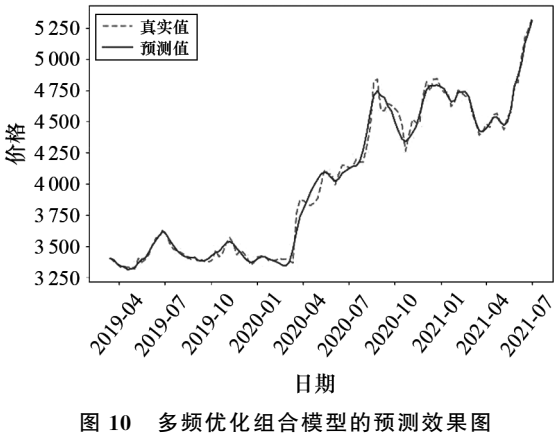

图10 多频优化组合模型的预测效果图

表 4　不同模型对大豆期货收盘价的预测结果

模型	RMSE	MAE	SMAPE
SVR	98.212	153.802	3.398
ARIMA	100.953	71.455	23.473
BPNN	137.203	98.590	2.083
LSTM	109.750	76.911	1.634
EMD-LSTM	105.932	78.519	1.700
CEEMDAN-BPNN（未重构）	66.729	50.404	1.091
CEEMDAN-LSTM（未重构）	97.031	74.090	1.554
多频优化组合模型	62.261	46.906	1.007

3　结论

利用本文构建的多频优化组合模型对我国大豆期货价格进行预测表明，该模型可以综合考虑影响大豆价格的多种因素，并依据不同单一模型的特性可充分提取重构后不同频率序列的波动特征，使得模型的预测精度显著优于其他单一模型以及 EMD-LSTM、CEEMDAN-BPNN（未重构）、CEEMDAN-LSTM（未重构）等组合模型，因此该模型可为大豆期货价格的预测提供良好参考. 在今后的研究中，我们将探讨其他单一模型在组合模型预测中的应用，以进一步提升模型的预测精度.

参考文献：

[1] 张婷. 基于 ARIMA 模型的国际粮食短期价格分析预测：以大豆为例[J]. 价格月刊，2016(7)：28-32.

[2] 张兵，刘丹. 不同类型主体对大豆期货价格波动的影响分析：基于向量自回归（VAR）模型[J]. 东南大学学报（哲学社会科学版），2012，14(6)：30-36.

[3] 柳苏芸，韩一军，包利民. 价格支持政策改革背景下国内外大豆市场动态关联分析：基于贝叶斯 DCC-GARCH 模型[J]. 农业技术经济，2016(8)：72-84.

[4] 何朋飞，李静，张冬青. APSO_SVR 模型在我国大豆价格预测的应用研究[J]. 大豆科学，2017，36(4)：632-638.

[5] 张冬青，刘欢，张云清. 基于 Q-RBF 神经网络模型的国产大豆价格预测研究[J]. 大豆科学，2017，36(1)：143-149.

[6] 范俊明，刘洪久，胡彦蓉. 基于 LSTM 深度学习的大豆期货价格预测[J]. 价格月刊，2021(2)：7-15.

[7] YU L，DAI W，TANG L. A novel decomposition ensemble model with extended extreme learning machine for crude oil price forecasting[J]. Engineering Applications of Artificial Intelligence，2016，47：110-121.

[8] CHAI J，WANG Y，WANG S Y，et al. A decomposition integration model with dynamic fuzzy reconstruction for crude oil price prediction and the implications for sustainable development[J]. Journal of Cleaner Production，2019，229：775-786.

[9] LI J M，WANG J. Stochastic recurrent wavelet neural network with EEMD method on energy price prediction[J]. Soft Computing，2020，24(22)：17133-17151.

[10] ZHOU J G，CHEN D F. Carbon price forecasting based on improved CEEMDAN and extreme learning machine optimized by sparrow search algorithm[J]. Sustainability，2021，13(9)：4896.

[11] 杨静凌，唐国强，张建文. 基于 CEEMD-Elman-Adaboost 组合模型的国际原油价格预测研究[J]. 重庆理工大学学报（自然科学），2021，35(3)：260-267.

[12] 高海翔，胡瑜，余乐安. 基于分解集成的 LSTM 神经网络模型的油价预测[J]. 计算机应用与软件，2021，38(10)：78-83.

[13] 方雪清，吴春胤，俞守华，等. 基于 EEMD-LSTM 的农产品价格短期预测模型研究[J]. 中国管理科学，2021，29(11)：68-77.

[14] 刘合兵，韩晶晶，庄晨辉，等. 基于多尺度特征融合的蔬菜价格预测模型研究[J/OL]. [2022-05-01] https://kns.cnki.net/kcms/detail/41.1112.S.20220325.1039.002.html.

[15] 贺毅岳，李萍，韩进博. 基于 CE 贺毅岳 EMDAN-LSTM 的股票市场指数预测建模研究[J]. 统计与信息论坛，2020，35(6)：34-45.