

基于复合降噪的 LSTM 综合模型的股指预测研究

杨亚男,叶善力

(浙江科技学院 理学院,杭州 310023)

摘要: 由于股指序列的预测难度大,为了改善预测精度,我们提出了一种基于复合降噪分解、优化预测和集成的综合研究方法来分析预测股票指数。首先,利用自适应噪声完备经验模态分解(complete ensemble empirical mode decomposition with adaptive noise,CEEMDAN)方法,将股票指数分解为不同频率的分量;然后,利用小波阈值(wavelet threshold,WT)去噪对合并的高频分量进一步降噪,得到新高频分量;最后再运用长短期记忆神经网络(long-short-term memory neural network,LSTM)模型分别对新高频分量、低频分量和趋势项进行预测,将预测结果求和即为股票指数的最终预测值。研究结果显示,CEEMDAN-WT-LSTM 在上证指数和深证成指上的平均绝对误差平均降低 32.22%,均方根误差平均降低 32.64%,决定系数平均提高 1.09%。CEEMDAN-WT-LSTM 与其他主流金融建模预测相比,其预测误差更低,具有更好的稳定性。本研究结果为股市股指预测提供了一种新方法。

关键词: 自适应噪声完备经验模态分解;小波阈值;长短期记忆神经网络;股指预测

中图分类号: F224.0 **文献标志码:** A **文章编号:** 1671-8798(2022)06-0465-09

Research on stock index prediction based on composite noise reduction LSTM comprehensive model

YANG Yanan, YE Shanli

(School of Sciences, Zhejiang University of Science and Technology, Hangzhou 310023, Zhejiang, China)

Abstract: Given the great difficulty in prediction of stock index series, a comprehensive research method was proposed on the basis of compound noise reduction decomposition, optimization prediction and integration to analyze and predict the stock index for the sake of improving the prediction accuracy. Firstly, the stock index was decomposed into components of different frequencies by using the method of complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN); secondly, wavelet threshold (WT) denoising was applied to further denoise the combined high-frequency components to obtain new high-frequency components; finally, the long-short-term memory neural network (LSTM) model was employed to predict the new high-frequency component, the low-frequency component and the trend item respectively, the sum of the prediction results being the final prediction value of the stock index. The research results show that the average absolute error of CEEMDAN-WT-

收稿日期: 2021-09-08

基金项目: 国家自然科学基金项目(11671357)

通信作者: 叶善力(1967—),男,福建省福州人,教授,博士,主要从事复函数空间、应用统计的金融数学研究。

E-mail: slye@zust.edu.cn。

LSTM on the Shanghai Stock Exchange Composite Index and Shenzhen Securities Component Index decreases by 32.22%, the average root mean square error decreases by 32.64%, and the determination coefficient increases by 1.09% on average. Compared with other mainstream financial modeling forecasts, CEEMDAN-WT-LSTM boasts lower prediction error and better stability. This study can provide a new method for the stock index prediction.

Keywords: CEEMDAN; wavelet threshold; LSTM; stock index prediction

由于股票市场的数据具有非线性且不稳定的特点,其数据预测比较困难。传统时间序列模型和机器学习方法是当前金融市场股票价格预测最常用的两类方法,其中,传统时间序列模型的预测方法主要有自回归移动平均模型、广义自回归条件异方差模型等^[1-2]。常月等^[3-4]证明传统组合模型比单一模型的预测精度更高、预测误差更低。如今,机器学习方法在金融领域备受关注且被广泛应用,主要包括支持向量机回归、卷积神经网络、循环神经网络(recurrent neural network, RNN)和长短期记忆神经网络(long-short-term memory neural network, LSTM)等^[5-7]。其中最受青睐的还是 LSTM 模型,杨青等^[8-9]验证了通过改变 LSTM 模型的参数可以达到比其他模型更优的预测精度。

随着人们对股市建模精度要求的提高,单一预测模型已经无法通过改进模型本身提高预测精度。1998 年, Huang 等^[10]提出经验模态分解(empirical mode decomposition, EMD),将其应用于处理非线性、非平稳时间序列的方法。为了改进 EMD 分解后模态混叠问题, Wu 等^[11-13]相继提出了集成经验模态分解(ensemble empirical mode decomposition, EEMD)、互补集成经验模态分解(complementary ensemble empirical mode decomposition, CEEMD)和自适应噪声完备经验模态分解(complete ensemble empirical mode decomposition with adaptive noise, CEEMDAN)等分解方法,这样既提高了分解效率,又降低了计算误差。在分解算法的基础上,姚洪刚等^[14-15]提出了 EMD-LSTM 和 CEEMDAN-LSTM 混合模型,证明其比单一 LSTM 模型预测误差更低。张培玲等^[16-17]针对单一分解方法,提出了 CEEMDAN-小波阈值(wavelet threshold, WT)和小波变换-EEMD 混合分解方法,证明其比单一分解方法去噪更优。为了进一步提高股票市场的预测精度,我们提出了基于 CEEMDAN-WT 去噪的复合降噪的 LSTM 综合预测模型。

1 相关模型与方法

1.1 EMD 模态分解

EMD 适合于非平稳信号的分析,将原始信号分解成不同本征模态函数(intrinsic mode function, IMF)。IMF 分量的确定需要满足以下两点要求:一是极值点要求,即至少包括一个极大值点和一个极小值点;二是局部最大和局部最小的上下包络线均值为零。EMD 的具体分解方法如下:

1) 对原始信号 $X(t)$ 确定局部最大值点和局部最小值点。

2) 利用 3 次样条函数对这些局部极大值点进行插值得到 $X(t)$ 的上包络线,同理得到局部极小值点下包络线,并求出上下包络线的均值

$$m(t) = \frac{u(t) + l(t)}{2} \quad (1)$$

式(1)中: $u(t)$ 为上包络线; $l(t)$ 为下包络线。

3) 分量 $h_1(t) = X(t) - m(t)$, 检查 $h_1(t)$ 是否满足 IMF 的条件,如果满足, $h_1(t)$ 则为第一个 IMF, 否则对 $h_1(t)$ 按照 $X(t)$ 做同样处理。重复上述迭代过程直到 $h_k(t)$ 符合 IMF 的条件或达到停止标准,停止标准的计算值为

$$y_{SD} = \sum_{t=0}^T \frac{|h_{k-1}(t) - h_k(t)|^2}{h_{k-1}^2(t)} \quad (2)$$

式(2)中: $h_{k-1}(t)$ 为第 $k-1$ 个 IMF 分量; $h_k(t)$ 为第 k 个 IMF 分量。当 y_{SD} 小于某一特定值即达到停止标准,其中特定值一般在 $[0.2, 0.3]$ 选取。

1.2 CEEMDAN 模态分解

EMD 虽然可以分解任意信号,但模态混叠,针对这个问题, Wu 等^[11]提出加入白噪声的 EEMD 改进方法。虽然模态混叠得到改善,但是出现了新的问题, EEMD 分解后会包含残余噪声,从而大大降低了重构信

号的预测精度。因此 Torres 等^[13]提出 CEEMDAN 方法,通过在原序列加入自适应白噪声,使 EEMD 方法得到了进一步的完善,提高了分解的速度,重构误差几乎趋向于零。CEEMDAN 具体算法步骤^[18]如下:

1) 产生含噪声的原序列:

$$S_i(t) = S(t) + \epsilon_0 \omega_i(t)。 \quad (3)$$

式(3)中: $S(t)$ 为原始时间序列; $\omega_i(t)$ 为服从正态分布 $N(0,1)$ 的高斯白噪声; ϵ_0 为噪声标准差; i 为添加噪声的次数。

2) 利用 EMD 对 $S_i(t) = S(t) + \epsilon_0 \omega_i(t)$ 进行分解,得到第一个 IMF:

$$y_{\text{IMF1}} = \frac{1}{T} \sum_{i=1}^T y_{\text{IMFi}}。 \quad (4)$$

式(4)中: y_{IMF1} 为 EMD 分解后第一个 IMF 分量; y_{IMFi} 为 EMD 分解后多个一阶分量。

3) 计算第一阶段残差:

$$r_1(t) = S(t) - y_{\text{IMF1}}。 \quad (5)$$

式(5)中: $r(t)$ 为残差; $r_1(t)$ 为第一阶段残差。

4) 对序列 $r_1(t) + \epsilon_1 E_1(\omega_i(t))$ 进行分解,得到第二个 IMF:

$$y_{\text{IMF2}} = \frac{1}{T} \sum_{i=1}^T E_1(r_1(t) + \epsilon_1 E_1(\omega_i(t)))。 \quad (6)$$

式(6)中: y_{IMF2} 为第二个 IMF 分量; $E_i(\cdot)$ 为 EMD 分解后第 i 个分量 IMF; ϵ_i 为噪声标准差。

5) 重复上述步骤,直到最后的残余分量不再进行分解,则为最后的残差,公式如下:

$$r(t) = S(t) - \sum_{j=1} y_{\text{IMFj}}。 \quad (7)$$

式(7)中: y_{IMFj} 为原始时间序列分解后得到的不同尺度分量。

1.3 小波阈值去噪

小波阈值去噪是通过小波对原始信号分解得到小波系数,采用恰当的阈值后,分别对系数和阈值的大小做对比处理及小波重构,最终达到降噪的目的。硬、软阈值函数是小波阈值去噪方法中最常见的两种阈值函数。硬阈值去噪方法是如果给定的阈值大于小波系数的绝对值,则为零;否则系数不变。硬阈值计算公式如下:

$$\omega_\lambda = \begin{cases} \omega, & |\omega| \geq \lambda; \\ 0, & |\omega| < \lambda. \end{cases} \quad (8)$$

软阈值去噪方法是如果给定阈值不大于小波系数的绝对值,则令其都减去阈值;否则,为零。软阈值计算公式如下:

$$\omega_\lambda = \begin{cases} [\text{sgn}(\omega)](|\omega| - \lambda), & |\omega| \geq \lambda; \\ 0, & |\omega| < \lambda. \end{cases} \quad (9)$$

式(8)~(9)中: ω 为小波系数; λ 为给定的阈值。

1.4 长短期记忆神经网络

LSTM 模型是一种特殊的 RNN 模型,不仅拥有了 RNN 模型的大部分特性,同时还解决了 RNN 梯度爆炸问题,即对远距离梯度形成依赖关系。LSTM 模型采用控制门的方式让信息选择性通过,而信息的更新和维持则需要依靠输入门、遗忘门和输出门来完成。LSTM 结构如图 1 所示。

x_t 是 LSTM 单元在的输入值, h_t 和 h_{t-1} 分别是 LSTM 在时刻 t 和时刻 $t-1$ 的输出值,而 c_t 是 LSTM 记忆单元时刻 t 的值。LSTM 模型具体算法步骤如下:

1) 计算遗忘门 f_t 的值。

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

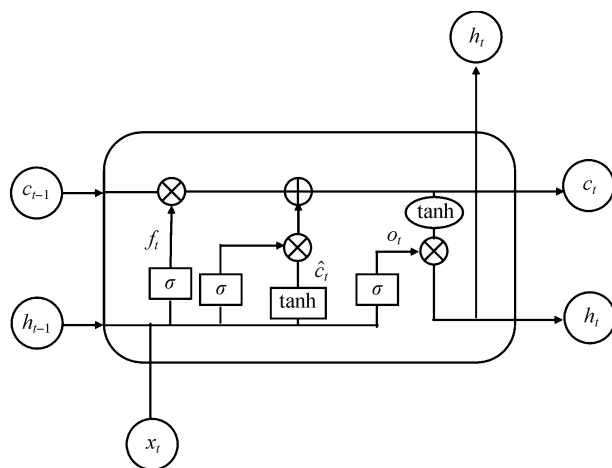


图1 LSTM 结构

Fig. 1 LSTM structure

式(10)中: \mathbf{W}_f 为遗忘门对应的权重矩阵; b_f 为遗忘门对应的偏置项。

2) 计算输入门 i_t 的值。

$$i_t = \sigma(\mathbf{W}_i \cdot [h_{t-1}, x_t] + b_i) \quad (11)$$

式(11)中: \mathbf{W}_i 为输入门对应的权重矩阵; b_i 为遗忘门对应的偏置项; σ 为 sigmoid 函数。

3) 计算 t 时刻候选记忆单元 \hat{c}_t 的值。

$$\hat{c}_t = \tanh(\mathbf{W}_c \cdot [h_{t-1}, x_t] + b_c) \quad (12)$$

式(12)中: \mathbf{W}_c 为候选记忆单元对应的权重矩阵; b_c 为候选记忆单元对应的偏置项; \tanh 为双曲正切函数,起激活作用。

4) 计算记忆单元 c_t 当前时刻的值。

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \hat{c}_t \quad (13)$$

式(13)中: c_{t-1} 为 LSTM 上一单元的状态值。

5) 计算输出门 o_t 的值。

$$o_t = \sigma(\mathbf{W}_o \cdot [h_{t-1}, x_t] + b_o) \quad (14)$$

式(14)中: \mathbf{W}_o 为输出门对应的权重矩阵; b_o 为输出门对应的偏置项。

6) 最后计算 LSTM 的输出值 h_t 。

$$h_t = o_t \times \tanh(c_t) \quad (15)$$

2 面向股市的建模思路

从上述分析我们可以看出 CEEMDAN 在数据分解上有较大的优势,而分解后的高频分量,可以利用小波阈值去噪进一步降噪,从而降低原序列的误差,并且 LSTM 在时间序列预测中对长期依赖性的学习能力非常突出。基于此,试验构建了 CEEMDAN 和小波阈值复合降噪的 LSTM 混合预测模型。CEEMDAN-WT-LSTM 建构流程如图 2 所示,其基本框架分为四个阶段:原始股指分解、高频分量优化、各序列预测和集成预测结果。具体步骤如下:

1) 基于 CEEMDAN 将原始股指 $S(t)$ 分解成 $res(residual, 趋势项)$ 与多个 IMF 分量;

2) 将所获的 IMF 通过单样本 t 检验的方法来确定高频分量和低频分量,将检验过的高频分量合并为一个新分量,记为 IMFS;

3) 利用小波阈值对高频分量 IMFS 去噪,得到新高频分量,记为 IMFH;

4) 利用 LSTM 模型分别对高频分量 IMFH、低频分量和趋势项进行建模预测,得到各分量的预测值,最后求和即为最终预测值。

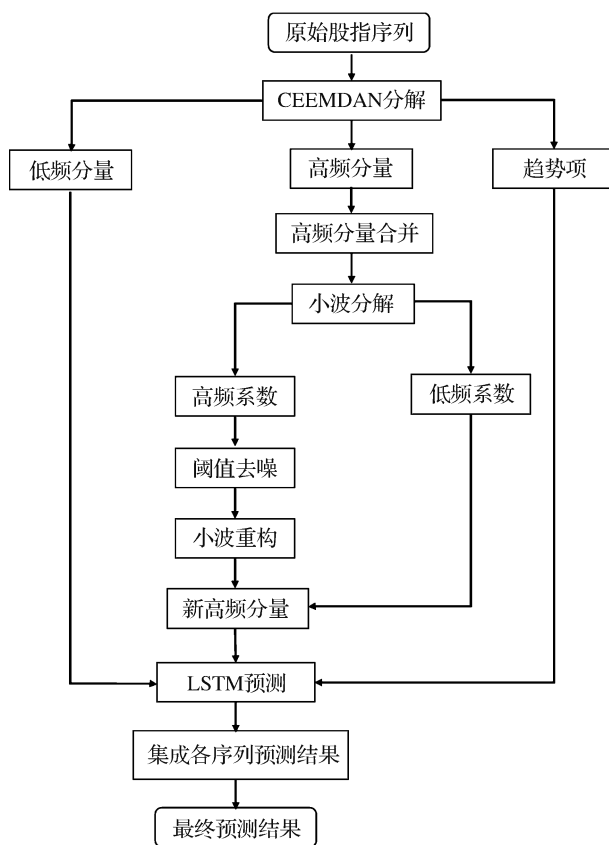


图 2 CEEMDAN-WT-LSTM 建构流程

Fig. 2 CEEMDAN-WT-LSTM construction process

3 试验分析

3.1 样本数据

选取上海证券交易所股价综合指数(Shanghai Stock Exchange Composite Index, SSEC, 简称上证指

数)、深圳证券交易所成份股价指数(Shenzhen securities component index, SZI, 简称深证成指)为数据建模基础,因上证指数和深证成指对投资者具有重要的参考价值,即通过对上证指数和深证成指的预测来判断股票市场变化的趋势。由于这两种股票指数的模型构建方法相似,故仅以上证指数为代表进行分析,深证成指不再具体介绍,只在最后试验结果中体现。

选取2015年1月5日至2021年5月31日之间上证指数的收盘价(来源为网易财经),对数据进行剔除空值,重新排序等预处理,共计1560个样本作为预测模型的原始数据。2020年10月20日之前共1410个数据作为训练集,2020年10月20日至2021年5月31日之间共150个数据作为测试集,以此分析预测的效果。

3.2 复合降噪分析

对原始股指序列添加20 dB的高斯白噪声,合成新的加噪股指序列。对此加噪合成股指序列分别采用小波阈值去噪和CEEMDAN-WT去噪这两种不同的算法进行去噪。这两种算法的优劣主要通过信噪比(signal-to-noise ratio, SNR)和均方误差(mean-square error, MSE)来进行标准评价,SNR和MSE的计算公式分别如下:

$$y_{\text{SNR}} = 10 \lg \left[\sum_{i=1}^n \frac{y^2(i)}{x^2(i) - y^2(i)} \right]; \quad (16)$$

$$y_{\text{MSE}} = \frac{1}{n} \sum_{i=1}^n \frac{y^2(i)}{x^2(i) - y^2(i)}。 \quad (17)$$

式(16)~(17)中: $x(i)$ 为去噪后的数据; $y(i)$ 为原始序列数据; n 为数据的长度。

根据SNR和MSE的评判标准,SNR值越大,MSE值越小,则评价结果越优。表1为SSEC在不同噪声强度下去噪性能指标数据,从数据对比可以看出,CEEMDAN-WT去噪的效果要优于单一小波阈值去噪。噪声强度从6 dB增加到20 dB,两者的信噪比都呈增加趋势,均方误差都呈下降趋势;但是对于同一噪声强度,CEEMDAN+WT去噪的信噪比明显高于小波阈值去噪,均方误差也明显小于小波阈值去噪。比如当噪声强度为20 dB时,CEEMDAN+WT去噪的信噪比和均方误差分别为21.091 5和0.051 3,比小波阈值去噪的信噪比提高了4.93%,比其均方误差下降了22.51%。因此,CEEMDAN-WT去噪算法的去噪效果更好一点。

表1 SSEC在不同噪声强度下去噪性能指标数据

Table 1 Denoising performance index data of SSEC at different noise intensities

噪声强度/dB	指标	小波阈值去噪	CEEMDAN+WT去噪
6	SNR/MSE	5.952 4/0.127 3	6.240 9/0.099 2
8	SNR/MSE	7.856 4/0.092 7	8.176 4/0.080 6
10	SNR/MSE	9.916 1/0.086 8	10.399 3/0.077 5
12	SNR/MSE	12.113 6/0.080 8	12.918 7/0.073 6
14	SNR/MSE	13.893 7/0.074 6	14.273 7/0.070 2
16	SNR/MSE	15.675 3/0.072 1	16.230 8/0.067 7
18	SNR/MSE	18.123 1/0.070 1	18.875 7/0.057 5
20	SNR/MSE	20.101 3/0.066 2	21.091 5/0.051 3

3.3 CEEMDAN分解

3.3.1 基于CEEMDAN的股票指数的序列分解

上证指数的CEEMDAN分解结果,如图3所示。上证指数的股票原始序列通过CEEMDAN算法可以分解为一个趋势项和7个IMF分量。将IMF1至IMF7从高频分量到低频分量依次排列,其波动的频率逐渐降低;对于高频分量,随着频率的降低,它对原始数据的影响逐渐增强。

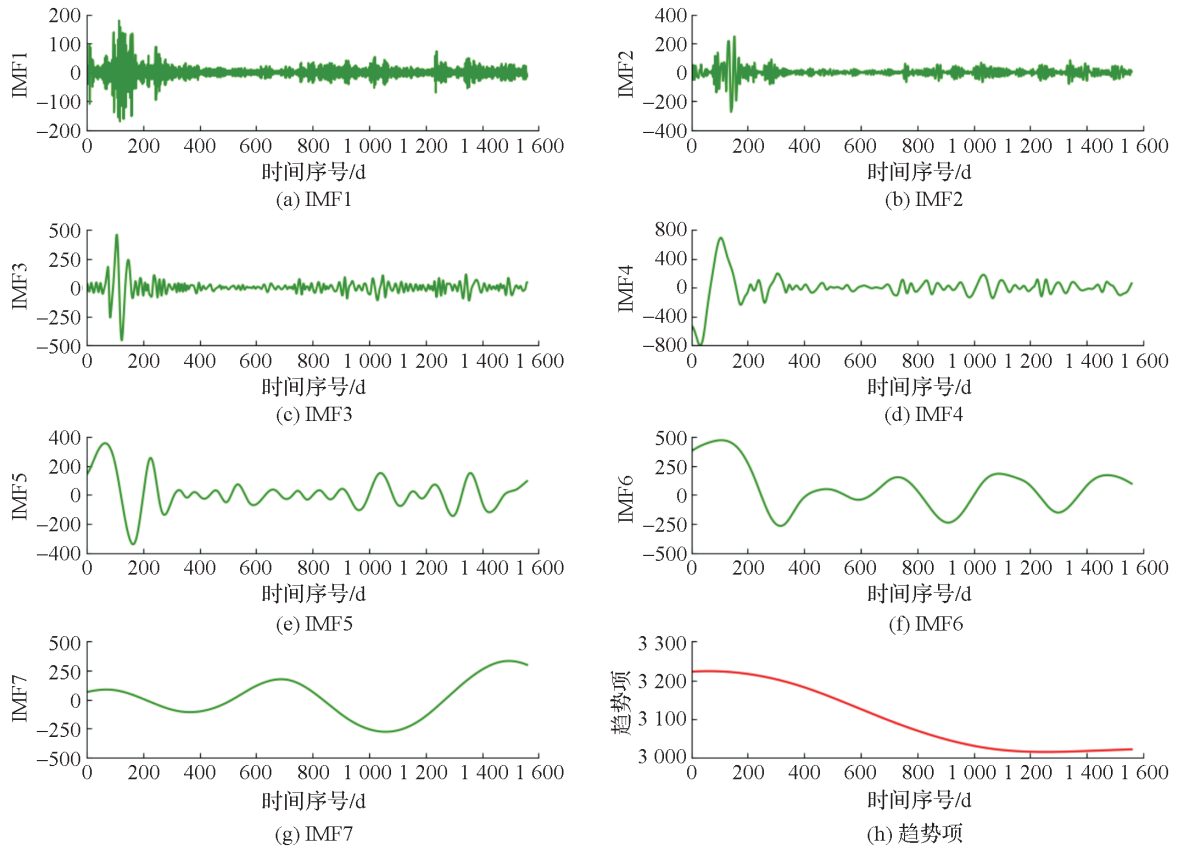


图 3 上证指数的 CEEMDAN 分解结果

Fig. 3 CEEMDAN decomposition results of Shanghai Stock Composite Exchange Index

3.3.2 高低频的确定

高频分量和低频分量的确定主要通过统计学中单样本 t 检验来判断。SSEC 训练集序列基于 CEEMDAN 分解的各分量单样本 t 检验见表 2。从表 2 中可知,单样本 t 检验,IMF1 的 t 检验值为 0.942,显著性(双尾)的数值为 0.346 大于 0.05,在 95%的置信区间下,不拒绝原假设,即 IMF1 均值与 0 没有差异。同理可以判断 IMF2、IMF3、IMF4 与 IMF1 一样,其显著性(双尾)的数值均大于 0.05,即均值与 0 没有差异,从而确定 IMF1 至 IMF4 均为高频分量,其主要反映随机波动对股票指数造成的影响。类比 IMF1 的分析过程,IMF5 的 t 检验值为 4.536,显著性(双尾)的数值为 0.00 小于 0.05,在 95%的置信区间下,拒绝原假设,即 IMF5 的均值与 0 存在显著差异;同理 IMF6 和 IMF7 与 IMF5 判断一样,其显著性(双尾)的数值均小于 0.05,即均值与 0 存在明显差异且都不相同,因此把 IMF5、IMF6、IMF7 视为低频分量,其主要体现不同的周期性波动对股票指数的影响。趋势项则主要代表股票指数的长期趋势。

表 2 SSEC 训练集序列基于 CEEMDAN 分解的各分量单样本 t 检验Table 2 Single sample t -test of each component based on CEEMDAN decomposition in SSEC training template sequence

分量	样本量	平均值	检验值=0					
			t	自由度	显著性(双尾)	平均值差值	差值 95%置信区间	
							下限	上限
IMF1	1 410	0.709	0.942	1 409	0.346	0.709	-0.768	2.185
IMF2	1 410	0.657	0.629	1 409	0.526	0.657	-1.391	2.704
IMF3	1 410	1.282	0.685	1 409	0.493	1.282	-2.388	4.951
IMF4	1 410	-3.396	0.686	1 409	0.493	-3.396	-13.109	6.317
IMF5	1 410	14.011	4.536	1 409	0.000	14.011	7.952	20.012
IMF6	1 410	53.219	10.338	1 409	0.000	53.219	43.119	63.317
IMF7	1 410	-20.706	-5.508	1 409	0.000	-20.706	-28.079	-13.332
res	1 410	3 108.061	1 459.805	1 409	0.000	3 108.061	3 103.885	3 112.238

3.4 小波阈值去噪

从图3上证指数CEEMDAN分解结果中可以看出高频分量仍然存有噪声,因此先对IMF1至IMF4进行求和得到高频分量之和IMFS,再利用小波阈值去噪对高频分量IMFS进行去噪处理。选取消失矩为3的Daubechies小波基函数,并采用3层分解结构对IMFS进行分解,选择启发式软阈值进行小波重构。小波阈值去噪前后的高频分量对比见图4。从图4可看出,高频分量经小波阈值去噪后,降噪效果较好。

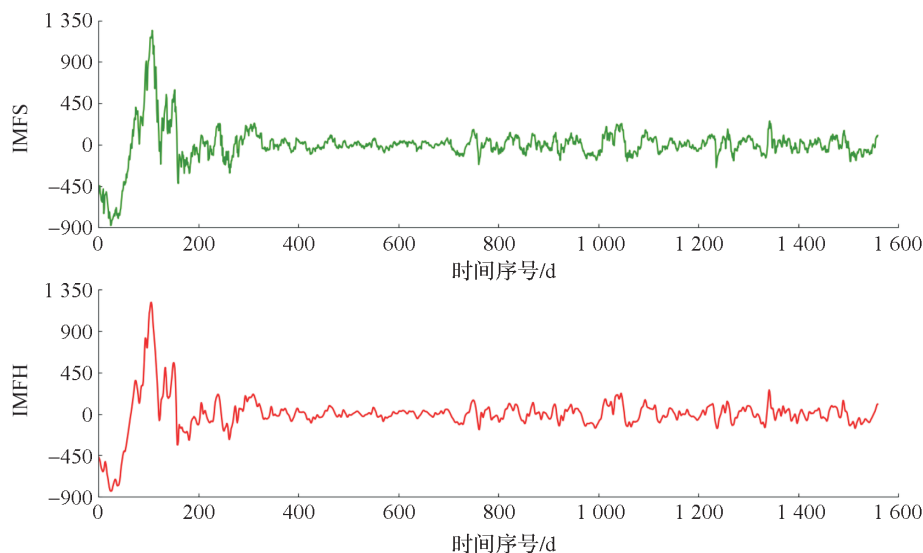


图4 小波阈值去噪前后的高频分量对比

Fig. 4 Comparison of high frequency sequence before and after wavelet threshold denoising

3.5 基于LSTM建模

正如前文所述,CEEMDAN将SSEC分解为4个高频分量,3个低频分量和一个趋势项,这4个高频分量合并后通过小波阈值去噪优化形成新分量IMFH;然后利用LSTM模型对各序列进行预测建模,主要采用两层LSTM结构层,其中还设置了一个随机失活层,防止数据出现过度拟合的问题。经过反复试验后,模型的参数设置最终确定为:每层神经元的个数为64,每轮训练样本的个数为64,迭代次数为500,损失函数为平均绝对误差,优化器为adam,时间步长为10。

模型搭建完成后,我们开始用训练好的模型进行预测,同时为了消除模型的偶然性,分别对各分量进行10次预测试验,并取10次预测结果的平均值,最终的试验结果即为各分量的预测结果,其计算公式如下:

$$\hat{y}_{SSEC} = \hat{y}_{IMFH} + \hat{y}_{IMF5} + \hat{y}_{IMF6} + \hat{y}_{IMF7} + \hat{y}_{res} \quad (18)$$

式(18)中: \hat{y}_{SSEC} 为上证指数最终预测值; \hat{y}_{IMFH} 为高频分量之和经过小波阈值去噪后的预测值; \hat{y}_{IMF5} 、 \hat{y}_{IMF6} 和 \hat{y}_{IMF7} 均为各低频分量预测值; \hat{y}_{res} 为趋势项的预测值。

3.6 评价指标

模型评价指标^[19]包括平均绝对误差(mean absolute error, MAE)、均方根误差(root mean square error, RMSE)及决定系数 R^2 ,计算公式分别如下:

$$y_{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|; \quad (19)$$

$$y_{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}; \quad (20)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (21)$$

式(19)~(21)中: \hat{y}_i 为预测值; y_i 为真实值; \bar{y} 为真实值的平均值。

由于深证成指的建模预测过程与上证指数的构建完全相似,因此只展示最终的预测结果。为了验证 CEEMDAN-WT-LSTM 模型的性能,将其与 LSTM 模型、EMD-LSTM 模型、CEEMDAN-LSTM 模型和 WT-LSTM 模型进行试验对比,观察不同模型对上证指数和深证成指的预测效果。各模型对 SSEC 和 SZI 的预测效果比较分别见表 3 和表 4,由表可知 CEEMDAN-WT-LSTM 模型在上证指数和深证成指上试验测试集上得到的 MAE 和 RMSE 均小于其他模型, R^2 分别为 0.996 2 和 0.988 3,更趋于 1;并且与其他模型对比后可知,CEEMDAN-WT-LSTM 模型的拟合效果更好,预测误差更低。

以上证指数为例,表现最优的 CEEMDAN-WT-LSTM 模型的 MAE 和 RMSE 值分别为 13.187 1 和 16.252 3;表现第二的为 CEEMDAN-LSTM 模型,MAE 值和 RMSE 值分别为 16.347 9 和 19.976 1。相比 CEEMDAN-LSTM 模型,CEEMDAN-WT-LSTM 模型的 MAE 和 RMSE 误差指标分别下降了 19.33%和 18.64%;与 EMD-LSTM 模型相比,该误差的下降幅度分别为 39.42%和 42.93%,改善效果明显;与 WT-LSTM 模型相比,该误差则分别下降了 48.48%和 48.65%,改善效果突出;与单一 LSTM 模型相比,该误差的下降幅度分别为 49.47%和 54.64%,改善效果颇为显著。这些数据证明原始数据经 CEEMDAN 分解和小波阈值复合降噪后的 LSTM 预测效果更佳,预测误差更小。

表 3 各模型对 SSEC 的预测效果比较

Table 3 Comparison of prediction effects of various models on SSEC

建模方法	R^2	MAE	RMSE
CEEMDAN-WT-LSTM	0.996 2	13.187 1	16.252 3
LSTM	0.981 3	26.095 8	35.826 8
EMD-LSTM	0.988 2	21.768 4	28.478 1
CEEMDAN-LSTM	0.994 2	16.347 9	19.976 1
WT-LSTM	0.985 4	25.596 5	31.649 7

表 4 各模型对 SZI 的预测效果比较

Table 4 Comparison of prediction effects of various models on SZI

建模方法	R^2	MAE	RMSE
CEEMDAN-WT-LSTM	0.988 3	58.476 3	72.473 5
LSTM	0.954 3	102.461 5	142.963 9
EMD-LSTM	0.981 7	77.475 9	90.583 9
CEEMDAN-LSTM	0.987 1	65.756 3	75.947 5
WT-LSTM	0.980 5	75.480 3	93.392 8

4 结 语

股票指数由于具有非稳定性、非线性的特点,传统预测模型已无法提高预测精度。本研究提出了 CEEMDAN 与小波阈值复合降噪的 LSTM 神经网络综合预测模型——CEEMDAN-WT-LSTM。首先通过引入 CEEMDAN 分解方法,提取序列波动信息;然后利用小波阈值去噪进一步降噪处理;最后再运用 LSTM 建立预测模型,求和即得最终预测结果。研究结果表明,在上证指数和深证成指的试验分析中,与其他模型相比,本研究提出的 CEEMDAN-WT-LSTM 模型对股票指数的预测具有更高的精度和稳定性,从而为金融时间序列数据预测提供了一种新方法。

参考文献:

- [1] 吴玉霞,温欣. 基于 ARIMA 模型的短期股票价格预测[J]. 统计与决策,2016(23):84.
- [2] 杨琦,曹显兵. 基于 ARIMA-GARCH 模型的股票价格分析与预测[J]. 数学的实践与认识,2016,46(6):81.
- [3] 常月,冯宇旭,曹显兵. 基于非线性时间序列模型的股票分析与预测[J]. 数学的实践与认识,2018,48(22):26.
- [4] 罗志丹,刘英,郭伟. 基于跟踪微分器的泰勒展开与 ARIMA 混合模型对股票价格的短期预测[J]. 2019,49(23):76.
- [5] 陈盟. 基于核主成分分析和支持向量回归对每日和每分钟股票价格的预测[D]. 兰州:兰州大学,2020.
- [6] 乔若羽. 基于神经网络的股票预测模型[J]. 运筹与管理,2019,28(10):134.
- [7] 冯宇旭,李裕梅. 基于 LSTM 神经网络的沪深 300 指数预测模型研究[J]. 数学的实践与认识,2019,49(7):309.
- [8] 杨青,王晨蔚. 基于深度学习 LSTM 神经网络的全球股票指数预测研究[J]. 统计研究,2019,36(3):76.
- [9] 彭燕,刘宇红,张荣芬. 基于 LSTM 的股票价格预测建模与分析[J]. 计算机工程与应用,2019,55(11):212.
- [10] HUANG N E, SHEN Z, LONG S R, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[J]. Proceedings of the Royal Society of London,1998,454(1971):921.

- [11] WU Z H, HUANG N E. Ensemble empirical mode decomposition: a noise-assisted data analysis method[J]. Advances in Adaptive Data Analysis, 2009, 1(1): 10.
- [12] YEH J R, SHIEH J S, HUANG N E. Complementary ensemble empirical mode decomposition: a novel noise enhanced data analysis method[J]. Advances in Adaptive Data Analysis, 2010, 2(2): 141.
- [13] TORRES M E, COLOMINAS M A, SCHLOTTHAUER G, et al. A complete ensemble empirical mode decomposition with adaptive noise[A]. Proceedings of 2011 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP 2011)[C]. Prague: IEEE, 2011: 4147.
- [14] 姚洪刚, 沐年国. EMD-LSTM模型对金融时间序列的预测[J]. 计算机工程与应用, 2021, 57(5): 243.
- [15] 贺毅岳, 李萍, 韩进博. 基于CEEMDAN-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): 44.
- [16] 张培玲, 李小真, 崔帅华. 基于改进小波阈值-CEEMDAN算法的ECG信号去噪研究[J]. 计算机工程与科学, 2020, 42(11): 2072.
- [17] 代军, 叶幸玮. 集合经验模式分解和小波变换方法的复合与应用[J]. 统计与决策, 2021, 37(13): 158.
- [18] 熊志斌. 基于CEEMDAN与LSTM的人民币汇率分析与预测[J]. 数据统计与管理, 2022, 41(3): 517.
- [19] 张倩玉, 严冬梅, 韩佳彤. 结合深度学习和分解算法的股票价格预测研究[J]. 计算机工程与应用, 2021, 57(5): 61.