DOI:10.13878/j.cnki.jnuist.20230903002

黄后菊[1]　李波[1]

基于 VMD-CSSA-LSTM 组合模型的股票价格预测

摘要

针对股票价格非平稳、非线性和高复杂等特性引发的预测难度大的问题,建立一种基于变分模态分解(Variational Mode Decomposition,VMD)-Circle 混沌映射的麻雀搜索算法(Circle Sparrow Search Algorithm,CSSA)-长短期记忆(Long Short-Term Memory,LSTM)神经网络的组合模型——VMD-CSSA-LSTM.首先,利用 VMD 将原始股票收盘价数据分解为若干本征模态函数(Intrinsic Mode Function,IMF)分量.然后,采用 Circle 混沌映射的 SSA 算法对 LSTM 神经网络的隐含层神经元、迭代次数、学习率进行优化,将最优参数拟合至 LSTM 网络中.最后,对每个 IMF 分量建模预测,将各分量预测结果叠加得到最终结果.实验结果表明,与其他模型相比,本文模型在多支股票数据集上的均方根误差(RMSE)、平均绝对误差(MAE)及平均绝对百分比误差(MAPE)均达到最小,预测股票收盘价格误差在 0 附近波动,稳定性更优、拟合更佳、精确度更高.

关键词

股票价格预测;变分模态分解;麻雀搜索算法;Circle 混沌映射;长短期记忆网络

中图分类号 TP391

文献标志码 A

收稿日期 2023-09-03

资助项目 辽宁省教育厅基本科研项目(面上项目)(JYTM20230862);国家自然科学基金(51679116);辽宁省自然科学基金(2020-MS-292)

作者简介

黄后菊,女,硕士生,研究领域为信号与信息处理.2807858629@qq.com

李波(通信作者),男,博士,教授,研究领域为智能信息处理.vincy1998@yeah.net

1 辽宁工业大学 电子与信息工程学院,锦州,121001

0 引言

股票预测是指通过分析股票市场的历史数据预测未来股票价格的变化趋势[1].在处理大规模、高复杂度的股票数据时,传统的统计预测方法难以获得理想的效果.随着信息技术的发展,深度学习凭借更专业的特征学习在股价预测过程中凸显其泛化性与准确性.

作为深度学习中典型的时序预测网络,长短期记忆(Long Short-Term Memory,LSTM)神经网络在股票价格领域较为表现出色,例如:Sun 等[2]利用 LSTM 网络提取数据中的时间特征,对上证指数(000001)进行了预测分析,预测效果优于传统的统计预测法;杨青等[3]利用深层 LSTM 网络对 30 只股票指数不同期限进行了预测研究,实验结果表明 LSTM 网络的泛化能力较为稳定.在利用 LSTM 网络预测股票的基础上,学者们提出了数据分解思想,典型分解有经验模态分解(Empirical Mode Decomposition,EMD)、动态模态分解(Dynamic Mode Decomposition,DMD)和变分模态分解(Variational Mode Decomposition,VMD).谢游宇等[4]构建的 EMD 与 LSTM 网络组合模型提高了预测精度,但容易发生模态混叠;史建楠等[5]利用 DMD-LSTM 混合模型对鞍钢股份进行了收盘价预测,虽能提取一定量模态信息,但所分解的模态过多,易混淆主辅模态;苏焕银等[6]构建了 VMD-LSTM 混合模型用于时变序列预测,证明了较 EMD、DMD 和单一 LSTM 网络,VMD-LSTM 能更好地拟合时变数据.此外,为提高预测精度,Zhang 等[7]由麻雀搜索算法(Sparrow Search Algorithm,SSA)确定了 LSTM 模型参数,相比经验论定义参数,SSA-LSTM 模型具有更高的预测精度,然而该算法存在初始化随机性的特点,在后期迭代过程中易陷入局部最优.

为了进一步提高预测精度、增强模型稳定性,针对复杂度高与非线性强的股票数据,本文提出一种融合 VMD、Circle 混沌映射的麻雀搜索算法(Circle Sparrow Search Algorithm,CSSA)与 LSTM 网络的股票价格预测模型——VMD-CSSA-LSTM.首先,利用 VMD 对原始股票收盘价序列进行变分模态分解,为将分解损耗约束到最低,使用约束条件确定分解模态数,得到 k 个表征局部特征的本征模态函数分量(Intrinsic Mode Function,IMF),在确保分解损耗为最低时剔除部分噪声分量,以此泛化非线性股票数据、降低数据复杂度.随后,利用 Circle 混沌映射初始化 SSA 算法,使得 SSA 初始化麻雀分布均匀,避免迭代

过程中陷入局部最优,并由该算法对 LSTM 的隐含层神经元、迭代次数、学习率参数寻优,以提高组合模型的鲁棒性.最后,将最优参数拟合到 LSTM 网络,对各 IMF 建模并预测股票收盘价,叠加各 IMF 与其余输入量预测结果得出最终预测值.

1 算法原理

1.1 VMD 分解算法

变分模态分解(VMD)是一种新型的时频分析方法,能把多分量信号一次分解为若干个单分量调幅调频信号,规避了迭代过程中的节点效应与虚假分量现象.VMD 方法利用构建并求解约束变分问题,将原始信号分解为特定数量的 IMF 分量,能有效处理非线性、非平稳信号.具体步骤如下:

1) 由希尔伯特变换求解出各模态的解析信号且构建频谱,得到每个模态函数在 t 时刻的解析信号:

$$s_k(t)=\left(\delta(t)+\frac{\mathrm{j}}{\pi t}\right)u_k(t). \tag{1}$$

式中:$u_k(t)$ 为第 k 个本征模态函数分量.

2) 对各模态解析信号估算的中心频率进行修正,将模态的频谱移到对应的基带.

3) 由解调信号的平方范数估算带宽,约束条件为带宽相加最小,其约束条件如下:

$$\begin{cases}\min\limits_{\{u_k\}\{\omega_k\}}\left\{\sum\limits_k\left\|\mu_t\left[\left(\delta(t)+\frac{\mathrm{j}}{\pi t}\right)u_k(t)\right]\mathrm{e}^{-\mathrm{j}\omega_k t}\right\|_2^2\right\},\\ \text{s.t.}\quad\sum\limits_k u_k(t)=y(t).\end{cases} \tag{2}$$

式中:ω_k 为第 k 个模态分量的中心频率[8];$\delta(t)$ 为单位脉冲函数;$y(t)$ 为初始信号.为将约束变分问题转换为非约束变分问题,本文引入二次惩罚因子 μ[9] 及拉格朗日乘子 λ,此处设 λ 初值为 0,$\langle,\rangle$ 为点积,其表达式如下:

$$\begin{aligned}L(\{u_k(t)\},\{\omega_k\},\lambda)=&\\ \mu\sum_K\left\|\mu_t\left[\left(\delta(t)+\frac{\mathrm{j}}{\pi t}u_k(t)\right]\mathrm{e}^{-\mathrm{j}\omega_k t}\right\|_2^2+&\\ \left\|y(t)-\sum_k u_k(t)\right\|_2^2+\langle\lambda(t),y(t)-\sum_k u_k(t)\rangle.&\end{aligned} \tag{3}$$

1.2 基于 Circle 混沌映射的麻雀搜索算法

1.2.1 麻雀搜索算法

麻雀搜索算法(SSA)主要模拟麻雀种群的捕食与反觅食的过程[10].该过程由发现者、加入者和预警者共同参与.发现者在种群中起到搜索和觅食作用,需要较高的适应度,搜索范围广.加入者主要追随发现者,适应度相对较低.预警者在察觉到种群中的捕食者时,对种群发出警告信息,发现者立即将种群迁徙到安全区域.于是,麻雀种群的矩阵表示为

$$\boldsymbol{X}=\begin{bmatrix}x_{11}&x_{12}&\cdots&x_{1d}\\x_{21}&x_{22}&\cdots&x_{2d}\\\vdots&\vdots&&\vdots\\x_{n1}&x_{n2}&\cdots&x_{nd}\end{bmatrix}. \tag{4}$$

式中:n 为麻雀种群的数量;d 为待优化问题变量的维度.

SSA 种群的适应度函数表示如下:

$$\boldsymbol{F}_x=\begin{bmatrix}f([x_{11}\quad x_{12}\quad\cdots\quad x_{1d}])\\f([x_{21}\quad x_{22}\quad\cdots\quad x_{2d}])\\\vdots\\f([x_{n1}\quad x_{n2}\quad\cdots\quad x_{nd}])\end{bmatrix}. \tag{5}$$

式中:f 为每只麻雀的适应度.

发现者的具体位置更新为

$$\boldsymbol{X}_{z,j}^{l+1}=\begin{cases}\boldsymbol{X}_{z,j}^{l}\times\exp\left(\dfrac{-z}{\alpha\times i_{\max}}\right), & R_2<T_{\mathrm{s}};\\ \boldsymbol{X}_{z,j}^{l}+Q\times\boldsymbol{L}, & R_2\geqslant T_{\mathrm{s}}.\end{cases} \tag{6}$$

式中:$l=1$ 为算法的迭代次数初值;$\boldsymbol{X}_{z,j}$ 表示第 z 只麻雀在第 j 维;$\boldsymbol{L}$ 为单位行向量;α 为[0,1] 间的随机数;$i_{\max}$ 为最终的迭代次数;Q 为标准正态分布随机数;R_2 为警告值且 $R_2\in[0,1]$;T_{s} 为安全值且 $T_{\mathrm{s}}\in[0.5,1]$.当 $R_2<T_{\mathrm{s}}$ 时,表示在该时刻种群范围内没有危险,发现者继续搜索,使种群有较高适应度;当 $R_2\geqslant T_{\mathrm{s}}$ 时,表示预警者察觉到危险信息,释放警告信号,种群向安全区靠近[11].于是,加入者追随发现者觅食过程表示为

$$\boldsymbol{X}_{z,j}^{l+1}=\begin{cases}Q\times\exp\left(\dfrac{\boldsymbol{X}_{\mathrm{worst}}^{l}-\boldsymbol{X}_{z,j}^{l}}{z^2}\right), & i>\dfrac{n}{2};\\ \boldsymbol{X}_{q}^{l+1}+|\boldsymbol{X}_{z,j}^{l}-\boldsymbol{X}_{q}^{l+1}|\times\boldsymbol{D}^{\mathrm{T}}(\boldsymbol{D}\boldsymbol{D}^{\mathrm{T}})^{-1}\times\boldsymbol{L}, & \text{其他.}\end{cases} \tag{7}$$

式中:$\boldsymbol{X}_{\mathrm{worst}}$ 为当下时刻种群最差位置;$\boldsymbol{X}_q$ 为当下时刻发现者的最佳位置;$\boldsymbol{D}$ 为 $1\times d$ 阶矩阵,其元素为 ±1 的随机值.通常,种群中有 10% ~ 20% 的麻雀作为预警者提供警告信息,其位置更新情况为

$$\boldsymbol{X}_{z,j}^{l+1}=\begin{cases}\boldsymbol{X}_{z,j}^{l}+K\times\left(\dfrac{|\boldsymbol{X}_{z,j}^{l}-\boldsymbol{X}_{\mathrm{worst}}^{l}|}{(g_z-g_{\mathrm{w}})+\varepsilon}\right), & g_z=g_{\mathrm{b}};\\ \boldsymbol{X}_{\mathrm{best}}^{l}+\beta\times|\boldsymbol{X}_{z,j}^{l}-\boldsymbol{X}_{\mathrm{best}}^{l}|, & g_z\neq g_{\mathrm{b}}.\end{cases} \tag{8}$$

式中:g_z 为当前时刻麻雀的适应度;g_{b} 为全局最优位

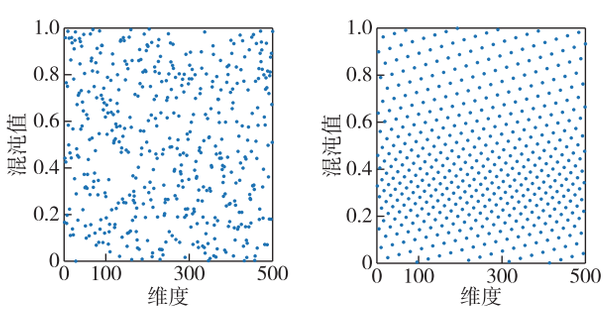
置的适应度;g_w 为全局最差位置适应度;X_{best} 为当前时刻的全局最优位置;β 是方差为1、均值为0的正态分布随机数,β 为控制步长的参数;K 为[-1,1]间的随机数;ε 为接近0的常数.当 $g_z \neq g_b$ 时,表示麻雀正处于种群边缘位置容易遭遇危险;当 $g_z = g_b$ 时,表示位于种群中心区域的麻雀收到危险信息,应向其他麻雀靠近以避免被捕食.

1.2.2 基于 Circle 混沌映射的麻雀搜索算法

SSA 可随机生成初始化种群,存在种群分布不均现象,致使中后期循环迭代种群多元性快速下降,陷入局部最优解难以跳出的问题.本文在初始化种群时采用 Circle 混沌映射改进种群分布情况,提升种群个体的多样化.$x_{i'}$ 为第 i' 个麻雀的位置,设 x_1 第一个麻雀的位置为随机初始化值,表达式如下:

$$x_{i'+1} = \mathrm{mod}\left[x_{i'} + 0.2 - \frac{1}{4\pi}\sin(2\pi x_{i'}),1\right]. \quad (9)$$

原始 SSA 随机初始化映射与 Circle 混沌映射对比如图 1 所示.可以看出,Circle 相对原始 SSA 随机映射种群分布更均匀,提高了个体的随机性.



a. SSA随机初始化　　b. Circle混沌映射初始化

图 1　两种种群初始化对比

Fig. 1　Comparison of two population initializations

1.3 LSTM 网络

LSTM 是一种改进的循环神经网络,采用 LSTM 能有效传送和表达较长时间序列中的信息且不会造成较长时间前的有效信息被遗忘.LSTM 网络的单元模块结构如图 2 所示.可以看出,LSTM 网络采用记忆细胞记录传递信息.LSTM 通过 3 个控制门来处理时滞任务,并利用 sigmoid 函数与 tanh 函数来更新单元状态.3 个门分别为遗忘门、输入门、输出门.遗忘门选择上一个阶段的信息多少能留存到现阶段单元状态,输入门选择现阶段输入信息多少能保存在现阶段单元状态,输出门选择现阶段单元状态有多少当作 LSTM 的输出值.

LSTM 单元状态计算公式为

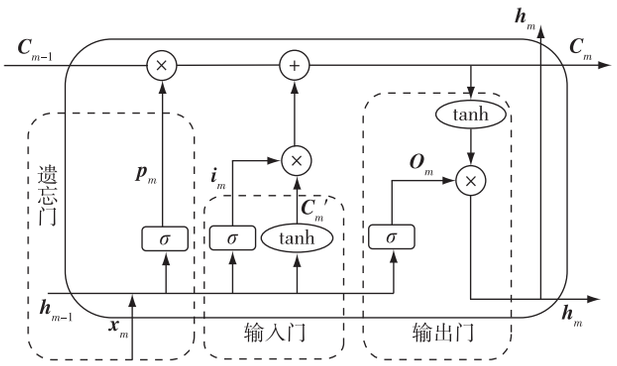


图 2　LSTM 结构

Fig. 2　LSTM structure

$$\boldsymbol{i}_m = \sigma(\boldsymbol{W}_i[\boldsymbol{h}_{m-1},\boldsymbol{x}_m] + \boldsymbol{b}_i), \quad (10)$$

$$\boldsymbol{C}'_m = \tanh(\boldsymbol{W}_c[\boldsymbol{h}_{m-1},\boldsymbol{x}_m] + \boldsymbol{b}_c), \quad (11)$$

$$\boldsymbol{P}_m = \sigma(\boldsymbol{W}_p[\boldsymbol{h}_{m-1},\boldsymbol{x}_m] + \boldsymbol{b}_p), \quad (12)$$

$$\boldsymbol{O}_m = \sigma(\boldsymbol{W}_o[\boldsymbol{h}_{m-1},\boldsymbol{x}_m] + \boldsymbol{b}_o), \quad (13)$$

$$\boldsymbol{C}_m = \boldsymbol{p}_m \odot \boldsymbol{C}_{m-1} + \boldsymbol{i}_m \odot \boldsymbol{C}'_m, \quad (14)$$

$$\boldsymbol{h}_m = \boldsymbol{O}_m \odot \tanh(\boldsymbol{C}_m). \quad (15)$$

式中:"⊙"为向量之间的点乘;σ 为 sigmoid 函数,其决定哪些信息将被更新;$\boldsymbol{i}_m$,$\boldsymbol{P}_m$ 和 $\boldsymbol{O}_m$ 分别为在第 m 个单元的输入、遗忘、输出门控;$\boldsymbol{x}_m$ 为第 m 个单元的输入;$\boldsymbol{h}_{m-1}$ 为第 $m-1$ 个单元的输出;$\boldsymbol{C}_m$ 为当前时刻的单元状态;$\boldsymbol{b}_i$,$\boldsymbol{b}_c$,$\boldsymbol{b}_p$ 和 $\boldsymbol{b}_o$ 分别为 $\boldsymbol{i}_m$,$\boldsymbol{C}'_m$,$\boldsymbol{P}_m$,$\boldsymbol{O}_m$ 的偏置项;$\boldsymbol{W}_i$,$\boldsymbol{W}_c$,$\boldsymbol{W}_p$ 和 $\boldsymbol{W}_o$ 分别为 $\boldsymbol{i}_m$,$\boldsymbol{C}'_m$,$\boldsymbol{P}_m$,$\boldsymbol{O}_m$ 的权重项;$\boldsymbol{C}'_m$ 为候选细胞信息,函数 tanh 用于创建新的 $\boldsymbol{C}'_m$,定义为

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \quad (16)$$

2 VMD-CSSA-LSTM 组合模型

本文提出的组合模型流程如图 3 所示,具体步骤表述如下:

1)组合模型共有 5 个输入量,分别为开盘价、最高价、最低价、成交量、收盘价.一个输出量为收盘价.本文利用 VMD 仅对输入量收盘价数据进行变分模态分解得到第 k 个 IMF 分量.

2)在 SSA 算法参数寻优前,利用 Circle 混沌映射初始化种群分布,并划分发现者、跟随者和预警者,设置迭代次数和参数上下边界.

3)为使 LSTM 网络结构和股票收盘价数据集最优匹配,由 SSA 算法在 LSTM 网络训练前对 LSTM 网络的隐含层神经元个数、迭代次数、学习率寻优.

4)数据集由收盘价、开盘价、最高价、最低价、

成交量 5 个维度构成.若设置 LSTM 网络步长为 n,则使用一个 n 行 5 列的矩阵数据对第 n+1 天的股票收盘价进行预测的计算量较大.于是,在每个分量建模后,应对数据集降维重构.

5) 将最优参数输入至 LSTM 模型,通过划分训练集测试集进行数据归一化.因 VMD 仅分解了输入量中收盘价数据,利用 VMD 分解后第 k 个收盘价数据的 IMF 分量与其余输入量共同预测收盘价,将 k 个 VMD 分解后的收盘价分量与其余输入量共同预测结果叠加得到最终某交易日的收盘价(如设网络步长 n 为 5,将前 5 个交易日的开盘价、最高价、最低价、成交量与 VMD 分解后收盘价的第 1 个 IMF 分量作为 CSSA-LSTM 模型输入,进行第 6 个交易日的收盘价预测,结果记为 A_1,继续将第 2 个 IMF 分量进行第 6 个交易日的收盘价预测,结果记为 A_2,直至得到 A_k,则第 6 日的最终预测结果为 $A_1+\cdots+A_k$),最后对各交易日收盘价预测数据进行反归一化,输出预测曲线.

3 实验结果与对比分析

3.1 数据来源

本文研究的原始数据集为 Tushare 财经共享数据集(http://tushare.org/),选取上证指数(000001)为主要实验股,选取 2013-02-22 至 2023-02-13 近 10 年的开盘价、收盘价、最高价、最低价、交易量为原始数据,共获取 2 425 个上证指数交易日历史数据(不含空数据),将 VMD 变分模态分解后的数据集作为样本数据集,选取样本数据集的前 70%,即 2013-02-22 至 2020-02-13 约 1 697 个交易日的数据作为训练样本集,样本数据集后 30%,即 2020-02-14 至 2023-02-13 约 728 个交易日的数据作为测试样本集.本文实验环境为 Windows 11、内存 16 GB 和 Matlab R2020b.

3.2 VMD-CSSA-LSTM 组合模型股票价格预测方法

3.2.1 VMD 变分模态分解

在 VMD 分解前,需确定所分解的 IMF 本征模态函数的数目,即确定 k 值.当 k 值过大时,邻近模态分量的中心频率较为贴近,会模糊掉部分信号,转换成一部分不需要的噪声分量,影响最终结果.当 k 值过小时,获得的 IMF 分量数目少于信号中有效成分数目,由于分解不完整,造成初始信号中某些关键信息被滤除.考虑到在分解过程中会生成无规律且变化幅度较大的残差,需要去除残差.由于去除的残差量会造成部分分解损失,本文针对预测精度与分解损失问题,定义分解损失约束条件并由 CSSA 优化算

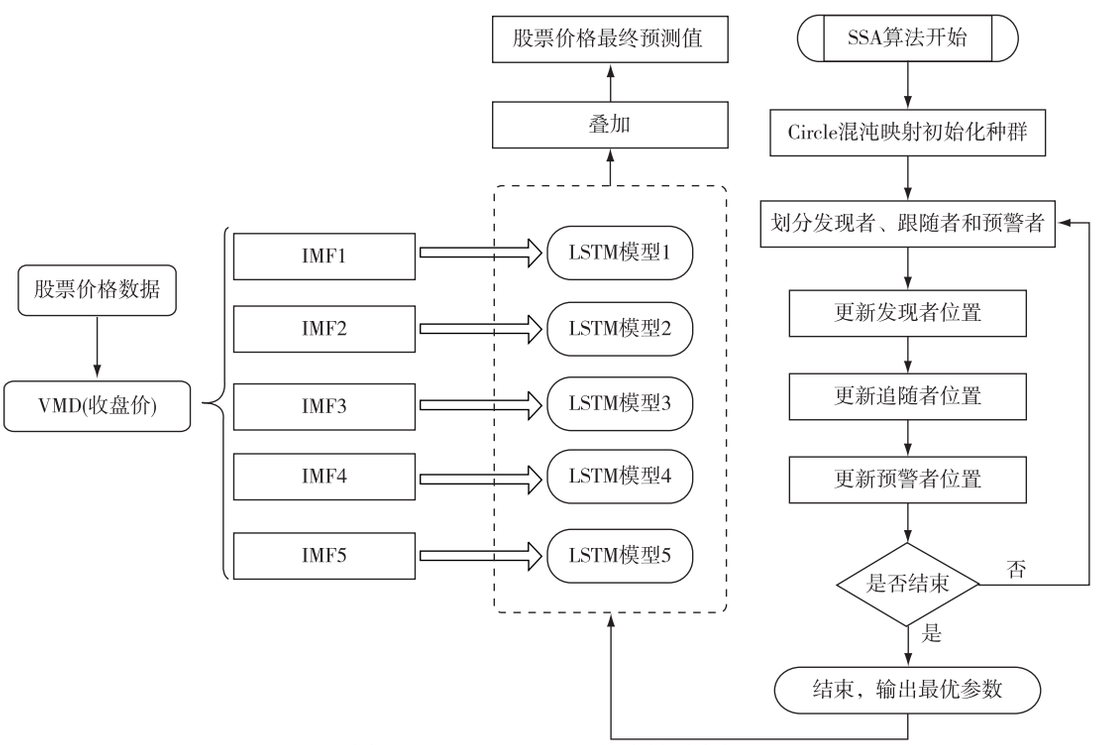


图 3 VMD-CSSA-LSTM 组合模型流程

Fig. 3 Flowchart of the proposed VMD-CSSA-LSTM combination model

法确定 $k(2 \leqslant k \leqslant 20)$ 值.设不考虑残差的序列为对应的本征模态函数分量之和:

$$y(t) = \sum_{k=2}^{20} u_k(t). \tag{17}$$

当考虑分解残差量 $R(t)$ 时,则残差的序列为

$$y'(t) = \sum_{k=2}^{20} u_k(t) + R(t). \tag{18}$$

于是,分解损失定义如下:

$$\min\left[\frac{1}{\bar{T}}\sum_{t=1}^{\bar{T}} \mid y(t) - y'(t) \mid\right] = \min\left[\frac{1}{\bar{T}}\sum_{t=1}^{\bar{T}} R(t)\right]. \tag{19}$$

式中:$\bar{T}$ 为序列的采样点数.可以看出,分解损失由 $R(t)$ 的平均值决定.为使分解损失达到最小值,此处将分解损失约束条件作为 CSSA 优化算法的目标函数,迭代次数设为 20,寻优 k 值.

图 4 通过约束条件确定 k 值,确保残差分量携带最少的有效信息并将其去除.实验得出,当 $k=5$ 时且迭代至第 17 次时,对上证指数(000001)股票收盘价的分解损失接近 1.65 且达到最低,确定 $k=5$ 为本实验的 IMF 分量数目.

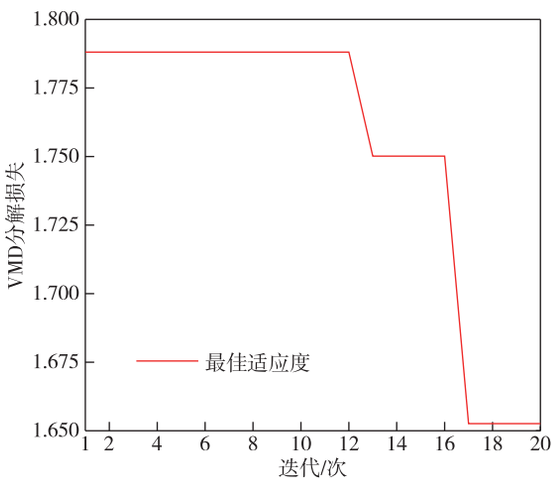


图 4　寻优 k 值进化收敛

Fig. 4　Evolutionary convergence of k-value seeking

在 VMD 算法中,输入序列为上证指数(000001)股票收盘价.其中:惩罚因子 $\mu=2\ 500$;噪声容忍度为 $\tau=0$,表示允许有误差;中心频率初始值为 1,表示中心频率均匀初始化;收敛精度为 10^{-6}.分解后的 5 个 IMF 分量如图 5 所示,本文将去除空数据后共计 2 425 个上证指数(000001)交易日历史收盘价数据分解为 5 个 IMF 分量,图中截取第 1~第 1 000 个交易日收盘价数据分解后的 IMF 分量.其中:第 1 行为原始序列信号;第 2~第 6 行分别为 VMD 分解的由低频到高频的 IMF1~IMF5 分量.可以看出,将上证指数(000001)收盘价数据分解成 5 个 IMF 分量,得到 5 个相对平稳的股票价格子序列.其中:IMF1 为频率最低的 IMF 分量,表示信号的走势或平均值;其他各分量表示原信号在各频段的波动变化,体现了信号的局部特征及其深层次信息.IMF5 体现了局部信号波动率的发展趋势,是最高频率分量.每个 IMF 分量既保留了原始股票价格信号的特征又避免了模态的混叠效应.

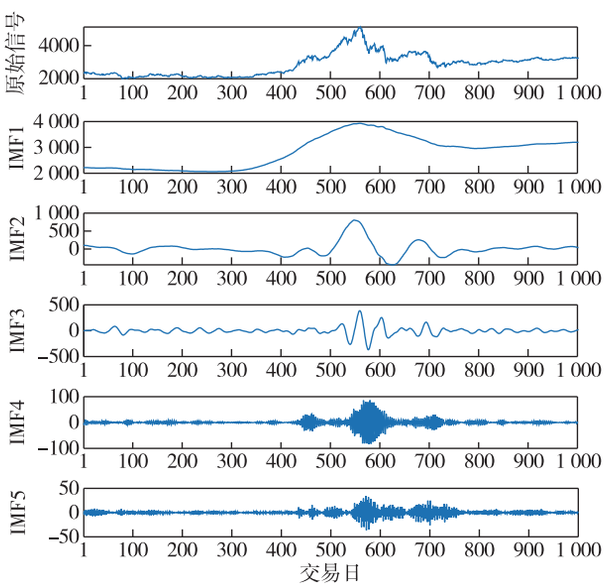


图 5　上证指数历史收盘价 VMD 分解

Fig. 5　Exploded view of VMD of the historical closing prices of the Shanghai Composite Index

3.2.2　CSSA-LSTM 组合模型算法

本文将分解后的 5 个 IMF 分量作为样本数据集输入 CSSA-LSTM 组合模型.为提高模型表达能力,使用激活函数加入非线性因素,该组合模型由输入层、LSTM 层、ReLU 激活层、输出层构成.为减少经验主观因素对组合模型的影响,使用 CSSA 算法对隐含层神经元数、迭代次数、学习率 3 个参数进行寻优.设置 LSTM 网络步长为 5,即以 5 个交易日的收盘价、开盘价、最高价、最低价、交易量预测第 6 天的上证指数(000001)收盘价,设置上下边界.

通常,隐含层神经元过少会导致模型欠拟合,过多会导致模型过拟合.当隐含层神经元个数小于 30 时预测结果欠拟合,而通过下边界从 1 开始依次乘自然数并代入模型,发现乘到 11 时的预测结果过拟合,因此确认上边界为 300.据观察多次改变隐含层神经元个数对预测结果的分析,隐含层神经元范围

在[30,300]间的效果较佳.

在对时间序列进行预测时,并非迭代次数越大,预测精度就越高.随着迭代次数增加,LSTM 网络中权重更新次数增加,预测结果会出现过拟合.据文献[12]可知,最大迭代次数在[0,400]间较优.为降低模型时间复杂度,本文通过对比迭代次数为 1,10,20,…,100 的预测结果拟合情况,确定下边界为 30.从 400 依次减 10 代入模型观察并预测结果拟合情况,最终确认迭代次数范围为[30,300].

学习率过大学习速度快,但 loss 容易出现梯度爆炸;学习率过小则收敛速度慢.现有研究常将学习率设为 0.1、0.01、0.001 和 0.0001 观察迭代损失情况.本文中将学习率为 0.1 时,损失值易震荡,学习率为 0.0001 时,收敛速度过慢,因此将学习率设为[0.001,0.01].

CSSA 算法对以上 3 个超参数寻优结果如表 1 所示.其中,最大迭代次数为 20,适应度函数选用均方根误差.考虑到 SSA 算法中预警者比例需占麻雀总数的 10%~20%可确保发现者向安全区移动,本文将预警者的比例设为最大比例 0.2.发现者在种群中起到搜索和觅食作用,需较高的种群数,因此将发现者的比例设为 0.7,跟随者的比例为 0.1,CSSA 算法同理.此外,模型训练过程优化器选用 Adam 算法.

表 1　CSSA 寻优 LSTM 网络超参数值

Table 1　Hyperparameter values of LSTM network by CSSA optimization

最优隐含层神经元数	最优迭代/次	最优学习率
260	258	0.009 8

在迭代过程中,VMD-CSSA-LSTM 组合模型在第 2 次收敛,均方根误差(Root Mean Square Error,RMSE)为 0.051 27,而经对比 VMD-SSA-LSTM 模型在第 4 次收敛,RMSE 误差为 0.056 53,本文 VMD-CSSA-LSTM 组合模型较 VMD-SSA-LSTM 模型收敛速度更快,RMSE 降低约 0.005 3.

接下来,将得到的最优超参数拟合至 LSTM 网络,对每个分量进行建模.为验证本文对于股票价格预测构建的 VMD-CSSA-LSTM 组合模型的可靠性和预测精度,对上证指数(000001)历史收盘价数据将 VMD 变分模态分解后的数据集作为样本数据集,选取该集合的前 70%,即 2013-02-22 至 2020-02-13 约 1 697 个交易日数据作为训练样本集,选取该集合的后 30%,即 2020-02-14 至 2023-02-13 约 728 个交易日数据作为测试样本集.本实验选取 LSTM、VMD-LSTM、VMD-SSA-LSTM 与本文组合模型对样本数据集进行预测对比,据文献[13],将 LSTM、VMD-LSTM 模型隐含层的神经元个数设为 100,学习率设为 0.001,最优迭代次数设为 10,VMD-SSA-LSTM 则通过 SSA 算法进行寻优获得隐含层的神经元个数为 272,学习率为 0.009 1,最优迭代次数为 275,再与本文模型做误差对比和预测结果对比.4 种模型真实值与预测值的差值对比如图 6 所示.可以看出,在对测试集共 728 条上证指数(000001)股票价格数据进行预测时,单一的 LSTM 网络对复杂非线性的上证指数(000001)价格预测误差波动在[-300,200]之间,本文模型的预测误差在 0 附近上下波动,误差较小,稳定性优越.

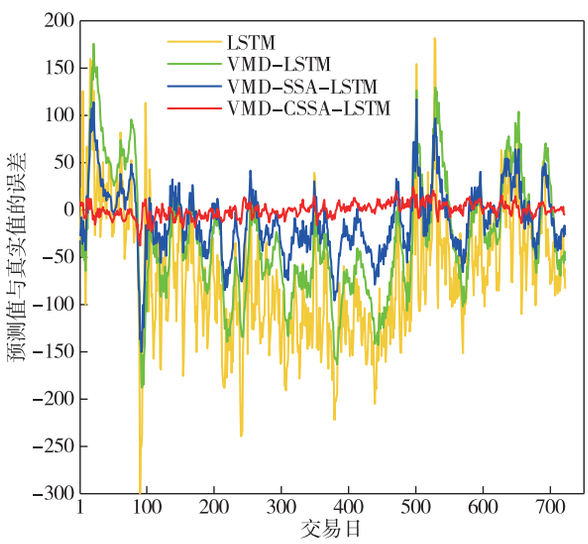


图 6　上证指数收盘价 4 种模型误差对比

Fig. 6　Comparison of errors at the closing prices of the Shanghai Composite Index predicted by 4 models

图 7 显示了上证指数(000001)股票收盘价预测结果.可以看出,上证指数(000001)股票收盘价在 2020-02-14 至 2023-02-13 约 728 个交易日内整体涨跌幅度波动较大,前期价格整体呈上涨阶段,收益趋势总体走强,本文提出的组合模型在整个过程表现最优,VMD-SSA-LSTM 表现次之.单一的 LSTM 网络与目标曲线拟合总体最差,特别是当股票价格涨跌幅波动较大时,本文模型能更为精准地贴合实际股票收盘价格,当股票收盘价涨到最高点时,VMD-CSSA-LSTM 组合模型曲线值最为接近目标预测值.在前期上涨阶段买入可获较大回报率,在测试集第 450 个交易日左右及时卖出,可有效避免严重的经

济损失.

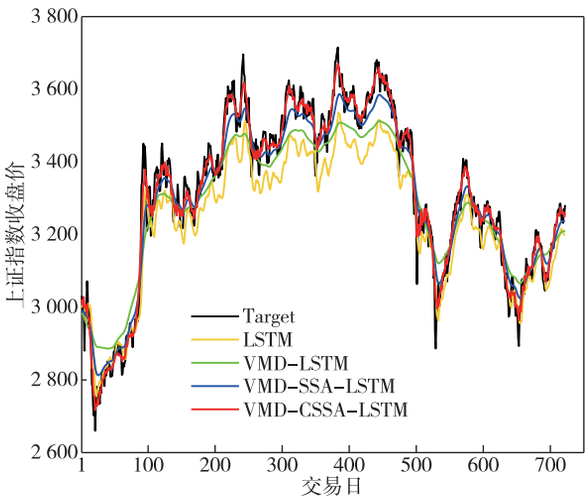


图 7　上证指数收盘价 4 种模型预测结果对比

Fig. 7　Comparison of the closing prices of the Shanghai Composite Index predicted by 4 models

为提高 3 个模型预测精确度与可信度,采用 RMSE、平均绝对误差(Mean Absolute Error,MAE)、平均绝对百分比误差(Mean Absolute Percentage Error,MAPE)3 种评价指标对模型进行性能指标评价,训练集性能指标如表 2 所示,测试集性能指标如表 3 所示.由表 2 可以看出:本文模型表现最优,该方法较单一的 LSTM 网络 RMSE 降低了 126.583 5,MAE 降低了 85.42,MAPE 降低了 3.007 4 个百分点;较 VMD-LSTM 预测方法 RMSE 降低了 108.041 6,MAE 降低了 65.400 1,MAPE 降低了 2.399 9 个百分点;较 VMD-SSA-LSTM 模型 RMSE 降低了 68.866,MAE 降低了 38.186 8,MAPE 降低了 1.440 9 个百分点.由表 3 可以看出,在测试集中本文模型同训练集表现最优.由此,在上证指数上的实验表明,在复杂的股票价格预测中本文模型更具优势,可以有效提高股票价格预测精度.

表 2　4 种模型的训练集预测性能指标

Table 2　Prediction performance indicators of the 4 models on training set

模型	RMSE	MAE	MAPE/%
LSTM	137.334 2	92.637 2	3.265 2
VMD-LSTM	118.792 3	72.617 3	2.657 7
VMD-SSA-LSTM	79.616 7	45.404 8	1.698 7
VMD-CSSA-LSTM	10.750 7	7.217 2	0.257 8

表 3　4 种模型的测试集性能指标

Table 3　Prediction performance indicators of the 4 models on test set

模型	RMSE	MAE	MAPE/%
LSTM	95.338 2	80.305 6	2.488 1
VMD-LSTM	64.617 8	53.639 2	1.624 4
VMD-SSA-LSTM	36.876 2	28.945 1	0.874 1
VMD-CSSA-LSTM	6.454 9	4.917 7	0.147 3

3.2.3　模型复杂度对比分析

本文模型复杂度主要由 LSTM 网络的时间复杂度和空间复杂度决定.其中,LSTM 网络的时间复杂度[14]计算公式如下:

$$O(W_{\text{time}}) = O(4MH^2 + 4MNH). \tag{20}$$

式中:M 为输入序列的长度;N 为输入特征的维度;H 为隐含层神经元数;W_{time} 为 LSTM 网络记忆单元计算量;$4MH^2$ 为输入门、遗忘门、输出门和候选记忆单元的计算量;$4MNH$ 为输入门、输出门和记忆单元的计算量.在降维重构后,存在 $M=1$ 和 $N=2\,425$. VMD 算法仅分解收盘价序列,在 VMD 算法将收盘价序列分解为 5 个分量后,将其维度由重构后的 M 值 1 变为 5,因此,其时间复杂度仅改变公式中的 M 值.由 3.2.2 节可知,在 LSTM 与 VMD-LSTM 模型中,H 值为 100.SSA 或 CSSA 算法在模型中起到参数寻优作用,其仅改变公式中的隐含层神经元个数 H 值.因此,在 VMD-SSA-LSTM 与 VMD-CSSA-LSTM 模型中,H 值为 SSA 或 CSSA 算法寻优后获得,分别为 272 和 260.

空间复杂度可描述算法所占内存空间,本文将其视为模型参数数量.因模型过万量级的参数量,本文计算空间复杂度时将 VMD、SSA 和 CSSA 自带个位数量级的参数忽略不计,LSTM 网络的空间复杂度计算公式如下:

$$O(W_{\text{space}}) = O(4HN + 4H^2 + 4H). \tag{21}$$

由表 4 可以看出,本文模型时间复杂度和空间复杂度均低于 VMD-SSA-LSTM 模型,高于 LSTM 和 VMD-LSTM 模型.由于 LSTM 网络本身的高复杂性和本文模型更高的预测精度,其时间复杂度和空间复杂度在可接受范围内.

表 4　4 种模型的复杂度对比

Table 4　Complexity comparison between the 4 models

模型	H	M	N	$O(W_{\text{time}})$	$O(W_{\text{space}})$
LSTM	100	1	2 425	$O(10.10\times10^5)$	$O(10.104\times10^5)$
VMD-LSTM	100	5	2 425	$O(50.50\times10^5)$	$O(10.104\times10^5)$
VMD-SSA-LSTM	272	5	2 425	$O(146.71\times10^5)$	$O(29.354\times10^5)$
VMD-CSSA-LSTM	260	5	2 425	$O(139.62\times10^5)$	$O(27.934\times10^5)$

3.2.4　模型有效性验证

为验证本模型的有效性和鲁棒性,本文另选取比亚迪(002594)、工商银行(IDCBY)和贵州茅台(600519)3只个股进行验证.实验环境、数据来源与模型同上证指数,选取2013-02-22至2023-02-13近10年的开盘价、收盘价、最高价、最低价、交易量为原始数据,划分前70%为训练集,后30%为验证集.当k=5时,由于工商银行(IDCBY)为美股,与另外2只个股略有差别,其分解损失约0.007美元.比亚迪(002594)分解损失约0.153元,贵州茅台(600519)分解损失约1.092元.由于美股节假日开盘情况与中国略微差别,因此工商银行(IDCBY)验证集为755个交易日数据.图8显示了这3只个股收盘价预测结果,红色曲线为本文模型,黑色曲线为实际目标数据,蓝色曲线为VMD-SSA-LSTM模型.可以看出,本文模型最为靠近真实目标曲线,VMD-SSA-LSTM模型稍次之,VMD-LSTM模型和LSTM网络虽与实际目标数据走势大致相同,但与实际目标数据拟合较差.本文模型较其他3个模型更为平稳,能反映股票的整体收盘价走势,且能在估计细微变化时取得较好的预测效果.

4　结语

为提高股票价格预测的稳定性和精确度,本文融合了VMD算法,引入Circle混沌映射的SSA算法和LSTM网络模型,提出一种组合模型——VMD-CSSA-LSTM.经过对上证指数(000001)收盘价预测的分析,得出如下结论:

1)VMD算法将股票价格时间序列分解成多个平稳的IMF分量,降低了数据复杂度,减少了部分测试误差噪声干扰,提高了预测精度.

2)在SSA算法引入Circle混沌映射初始化种群分布,提高了迭代收敛速度并降低了迭代误差.选择CSSA算法进行LSTM网络隐含层参数优化,提高了预测的稳定性和鲁棒性.

由于股票收盘价格预测在一定程度上受投资者主观因素影响,本文模型还有改进的空间.下一步拟将投资者情绪引入到本文模型以取得更好的预测结果.此外,本文模型侧重于预测精度的提高,其时间复杂度和空间复杂度难以兼顾到最低.因此,对本文模型时间复杂度和空间复杂度的优化还有待深入研究.

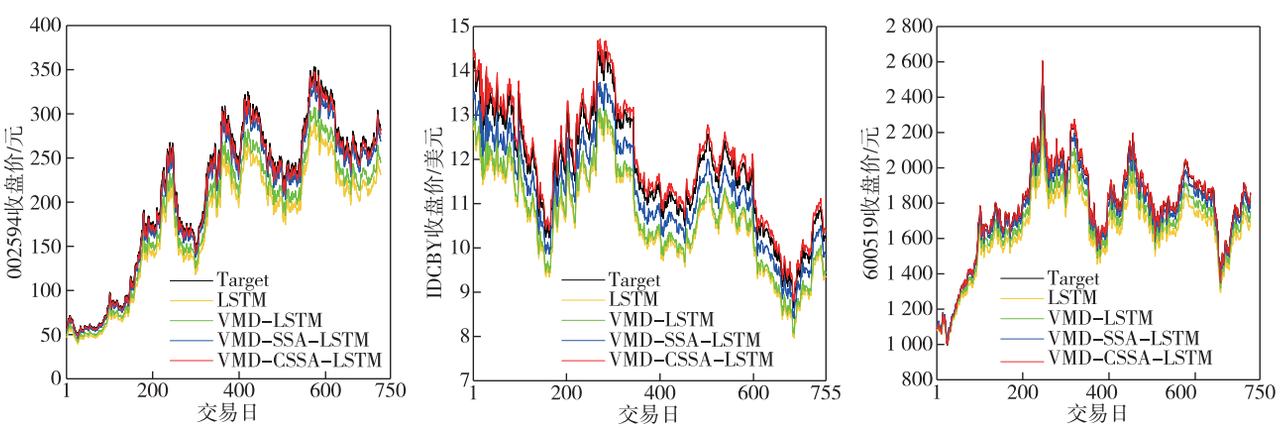


图8　3只个股收盘价4种模型预测结果对比

Fig. 8　Comparison of closing prices of 3 stocks predicted by the 4 models

参考文献

References

[1] 杨智勇,叶玉玺,周瑜.基于BiLSTM-SA-TCN时间序列模型在股票预测中的应用[J].南京信息工程大学学报(自然科学版),2023,15(6):643-651
YANG Zhiyong, YE Yuxi, ZHOU Yu. Application of BiLSTM-SA-TCN time series model in stock price prediction [J]. Journal of Nanjing University of Information Science & Technology (Natural Science Edition), 2023, 15(6): 643-651

[2] Sun Y, Sun Q S, Zhu S. Prediction of Shanghai stock index based on investor sentiment and CNN-LSTM model [J]. Journal of Systems Science and Information, 2022, 10(6): 620-632

[3] 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):65-77
YANG Qing, WANG Chenwei. A study on forecast of global stock indices based on deep LSTM neural network [J]. Statistical Research, 2019, 36(3): 65-77

[4] 谢游宇,王万雄.基于EMD和SSA的股票预测模型[J].计算机工程与应用,2023,59(18):285-292
XIE Youyu, WANG Wanxiong. Stock forecasting model

based on EMD and SSA[J].Computer Engineering and Applications,2023,59(18):285-292
[5] 史建楠,邹俊忠,张见,等.基于 DMD-LSTM 模型的股票价格时间序列预测研究[J].计算机应用研究,2020,37(3):662-666
SHI Jiannan, ZOU Junzhong, ZHANG Jian, et al. Research of stock price prediction based on DMD-LSTM model[J].Application Research of Computers,2020,37(3):662-666
[6] 苏焕银,彭舒婷,曾琼芳,等.基于 VMD-LSTM 混合模型的城际高速铁路时变客流预测[J].铁道科学与工程学报,2023,20(4):1200-1210
SU Huanyin, PENG Shuting, ZENG Qiongfang, et al. Forecast of time-dependent passenger flow of intercity high-speed railway based on VMD-LSTM mixed model [J].Journal of Railway Science and Engineering,2023,20(4):1200-1210
[7] Zhang Y Y,He D,Wu Q Y.Forecasting of PM2.5 concentration time series based on SSA-LSTM model[C]// International Conference on Statistics,Data Science,and Computational Intelligence (CSDSCI 2022).SPIE,2023,12510:373-380
[8] 姜超,李国富.改进 VMD-LSTM 法在刀具磨损状态识别中的应用[J].机械科学与技术,2022,41(2):246-252
JIANG Chao,LI Guofu.Application of modified VMD and LSTM in tool wear state recognition model [J]. Mechanical Science and Technology for Aerospace Engineering,2022,41(2):246-252
[9] 张晨阳,张亚,李培英,等.基于变分模态分解的侵彻过载信号特征提取[J].探测与控制学报,2021,43(3):16-21
ZHANG Chenyang,ZHANG Ya,LI Peiying,et al.Feature extraction of penetration overload signal based on variational mode decomposition[J].Journal of Detection & Control,2021,43(3):16-21
[10] 左亚辉,谢源,邹定江,等.基于混沌麻雀搜索算法的 PMSM 直接转矩控制[J].组合机床与自动化加工技术,2023(2):174-177
ZUO Yahui, XIE Yuan, ZOU Dingjiang, et al. PMSM direct torque control based on chaotic sparrow search algorithm[J].Modular Machine Tool & Automatic Manufacturing Technique,2023(2):174-177
[11] 柴岩,孙笑笑,任生.融合多向学习的混沌麻雀搜索算法[J].计算机工程与应用,2023,59(6):81-91
CHAI Yan,SUN Xiaoxiao,REN Sheng.Chaotic sparrow search algorithm based on multi-directional learning[J]. Computer Engineering and Applications,2023,59(6):81-91
[12] 刘明,宁静.基于 SSA-LSTM 的重大突发疫情演化预测方法[J].信息与管理研究,2022,7(6):16-29
LIU Ming,NING Jing.Evolutionary prediction method of major epidemic outbreak based on SSA-LSTM [J]. Journal of Information and Management,2022,7(6):16-29
[13] 李秀昊,刘怀西,张智勇,等.基于 VMD-LSTM 的超短期风向多步预测[J].南方能源建设,2023,10(1):29-38
LI Xiuhao, LIU Huaixi, ZHANG Zhiyong, et al. Very short-term wind direction multistep forecast based on VMD-LSTM[J].Southern Energy Construction,2023,10(1):29-38
[14] Lin M L,Chen C X.Short-term prediction of stock market price based on GA optimization LSTM neurons[C]// Proceedings of the 2018 2nd International Conference on Deep Learning Technologies,2018:66-70

Stock price prediction based on VMD-CSSA-LSTM combination model

HUANG Houju[1] LI Bo[1]
1 School of Electronics & Information Engineering,Liaoning University of Technology,Jinzhou 121001,China

Abstract To address the problems of stock price prediction due to its non-static,highly complex and random fluctuations,a combination model based on Variational Mode Decomposition (VMD)-Circle Sparrow Search Algorithm (CSSA)-Long Short-Term Memory (LSTM) neural network is established.The original stock closing data is decomposed into several Intrinsic Mode Function (IMF) components by VMD,and then the CSSA is used to optimize the parameters of hidden layer neurons,iteration number and learning rate of LSTM,and the optimal parameters are fitted into the LSTM,where each IMF component is modeled and predicted,and the prediction results of IMF component are superimposed to obtain the final result.Experiments show that the RMSE,MAE and MAPE of the proposed model are minimized on multiple stock datasets,the error of the predictied closing prices of individual stocks fluctuates around 0,which is more stable with better fitting and higher accuracy.

Key words stock price forecasting;variational mode decomposition (VMD);sparrow search algorithm (SSA);Circle chaos mapping;long short-term memory (LSTM)