基于优化 VMD 再分解的 LSTM 下的股价预测

王艺涵[1] 路 翀[1] 龙 洁[1] 雷一鸣[1]
WANG Yihan LU Chong LONG Jie LEI Yiming

摘 要 研究设计了一种股价预测混合模型，以应对股价受多种因素影响的挑战。模型结合了完备集合经验模态分解（CEEMDAN）与经粒子群优化（PSO）的变分模态分解（VMD）技术，并使用长短期记忆网络（LSTM）对股价进行预测，旨在处理股票价格中的噪声并提升预测精度。通过与 LSTM 模型、循环神经网络（RNN）模型、卷积神经网络（CNN）模型和支持向量回归（SVR）模型的对比，实验结果表明，所设计的模型相较于其他模型表现出更高的鲁棒性和准确性，对于金融从业者在制定投资策略时具有指导意义，同时有助于深度学习在股价预测领域的应用，具备实际应用价值。

关键词 股价预测；混合模型；粒子群优化算法；完备集合经验模态分解；长短期记忆网络

doi：10.3969/j.issn.1672-9528.2024.07.001

0 引言

在金融领域，尤其是股价预测这一分支，众多股价预测算法已被广泛应用于学术研究中 [1-3]。目前的各种预测模型主要有三种：单一的统计模型、智能模型和混合模型。

（1）在各种预测模型的初期，主要采用的是单一的统计模型来预测股价，统计模型可以通过统计历史数据对未来进行预测分析，例如 Prasad 等人 [4] 对比了卡尔曼滤波器（Kalman filter）、分布式梯度增强库（XGBoost）和自回归移动平均模型（ARIMA）这三种主要的算法在股价预测中的应用，Kalman filter 由于其递归和反馈机制，被认为适合精确预测。XGBoost 适合非线性数据集，并能有效地捕捉时间依赖特征。ARIMA 模型则在时间序列数据上广受欢迎，并通过消除数据的平稳性来工作。统计模型往往假设变量之间的关系是线性且稳定的，但是股价的波动特征复杂，受多方面因素的影响，统计模型往往无法很好地处理这些数据。

（2）相较单一的统计模型，智能模型在处理复杂的非线性和非平稳数据方面显示出更加强大的能力。例如，Van Gestel 等人 [5] 利用最小二乘支持向量机（LSSVM）对金融时间序列进行分析，Lee[6] 使用支持向量机（SVM）对股票趋势进行预测，而 Qiu 等人 [7] 深入探讨了长短期记忆网络（LSTM）的内存单元结构和门控机制，以及集成注意力机制如何提升模型在处理和预测时间序列数据，尤其是股市预测中的表现。这些研究表明，智能模型能够取得令人瞩目的成果。然而，这些模型在应用条件上的局限性有时会影响预测的准确性。

（3）近几年，混合模型的出现，极大提升了对复杂数据预测的准确性。例如，Jing 等人 [8] 使用卷积神经网络 (CNN) 对股票投资者的隐藏情绪进行分类，然后结合 LSTM 对股票的市场价格进行分析预测，取得了比基线分类器更好的性能。Yujun 等人 [9]，利用经验模态分解（EMD）将复杂的原始股票价格时间序列分解为多个子序列，然后结合 LSTM 的方法来训练和预测每个子序列，最终将子序列进行融合得到最终的预测，实验结果表明混合预测的方法具有实际应用价值。Pai 等人 [10] 提出一种使用 ARIMA 和 SVM 的混合模型来预测股票价格，实验结果证明，这种模型在股票预测方面稳定性较好，但是因为计算量庞大，而导致运行时间比较长。

如今，各种混合模型的预测方法层出不穷，基于分解积分方法进行预测的模型更是应用在了各大领域 [11-13]，已经可以证明其有效性和稳定性。本研究主要基于分解积分方法提出一种全新的混合模型来对股票收盘价格进行预测。使用完备集合经验模态分解（CEEMDAN）和变分模态分解（VMD）用于噪声的抑制，使用 LSTM 进行股价预测，利用样本熵（SE）用来减少计算时间，利用粒子群优化算法（PSO）避免 VMD 参数选择的主观性，更好地抑制噪声。

本研究主要是根据前一天股票的收盘价来预测第二天股价，主要的创新体现在以下方面：（1）本研究利用 PSO 与 VMD 相结合，避免 VMD 参数的主观选择，以提高模型的预测能力；（2）本研究提出一种全新的 CEEMDAN-SE-PSOVMD- 混合 LSTM 的混合模型并在不同数据集运行，验证了模型预测的准确性和鲁棒性。

1. 新疆财经大学信息管理学院 新疆乌鲁木齐 830000
[基金项目] 国家自然科学基金（No.62166039）

1 基本理论

1.1 CEEMDAN

CEEMDAN[14] 是由 EMD 发展而来，其本质是基于 EMD 可以自适应地将原始数据分解为多个本征模态函数 (IMFs)，在股票价格预测等具有非线性和非平稳的时间序列数据上具有极强的适应性。CEEMDAN 的具体步骤如下。

设 $f(t)$ 为原始数据，$IMF_k(t)$ 为 CEEMDAN 方法得到的第 k 个 IMF。$EMD_j(\cdot)$ 表示 EMD 分解得到的第 j 个 IMF。ε_k 用于设置每一级的信噪比（SNR），它决定了白噪声的标准差。$\omega^i(t)$ 是满足标准正态分布的高斯白噪声。在这部分的计算中，除了 ε_k 是标量系数外，其他变量都代表一个序列长向量，包括 $f(t)$、$IMF_k(t)$、$\omega(t)$、$r(t)$。

（1）在原始时间序列数据 $f(t)$ 中添加 SNR 为 ε_0 的白噪声 $\omega^i(t)$（i=1,2,3,⋯,n），得到数据 $f^i(t)$（i=1,2,3,⋯,n）用于第一次分解，如式（1），其中 t 代表不同的时间点，i 代表添加的第 i 个白噪声，n 为总数添加白噪声的次数。股票价格的时间序列被假设为随机信号，其幅度不可预测，但遵循特定的统计特征。

$$f^i(t) = f(t) + \varepsilon_0\omega^i(t), i = 1,2,3, \ldots., n \tag{1}$$

（2）使用 EMD 将 $f^i(t)$ 分解 n 次，得到 n 个第一 IMF 结果，命名为 $MF^i_1(t)$（t=1,2,3,⋯,n）。然后利用公式（2）计算均值得到 CEEMDAN 的第一个 IMF，$IMF_1(t)$，利用公式（3）求出第一个残差 $r_1(t)$，其中 $EMD_1(\cdot)$ 代表由 EMD 获得的第一个 IMF。理论上，由于白噪声的平均值为 0，可以通过计算平均值来消除白噪声的影响。

$$\overline{IMF_1}(t) = \frac{1}{n}\sum_{i=1}^{n} IMF_1^i(t) = \frac{1}{n}EMD_1\left(f^i(t)\right) \tag{2}$$

$$r_1(t) = f(t) - \overline{IMF_1}(t) \tag{3}$$

（3）将自适应噪声项添加到第一个残差 $r_1(t)$ 中，并获得新的时间序列。自适应噪声项是由白噪声 $w^i(t)$ 产生的，其强度可以根据信噪比（SNR）进行调整，表示为 $\mathrm{SNR}_{\varepsilon 1}$，对这个加入了自适应噪声的信号进行 EMD 分解，以得到第一个 IMF。然后，以新的为载体，进行 EMD 分解，即可得到 CEEMDAN 的第二个 IMF，$\overline{IMF_2}(t)$，如式（4）所示。同时，可得到第二个余数 $r_2(t)$，如式（5），其中 $EMD_1(\cdot)$ 表示 EMD 得到的第一个 IMF。

$$IMF_2(t) = \frac{1}{n}\sum EMD_1\left(r_1(t) + \varepsilon_1 EMD_1\left(\omega^i(t)\right)\right) \tag{4}$$

$$r_2(t) = r_1(t) - \overline{IMF_2}(t) \tag{5}$$

（4）将新的自适应噪声项添加到具有白噪声的 $\omega^i(t)$（i=1,2,3,⋯,n）和 $SNR\varepsilon_k$（k=2,3,⋯,K）中，其中 K 是完成 CEEMDAN 后的 IMF 总数。然后，如式（6）可得到 CEEMDAN 的第 k 个 IMF，$IMF_k(t)$。如式（7）可得到第 k 个残差 $r_k(t)$，其中 $EMD_k(\cdot)$ 表示 EMD 得到的第 k 个 IMF。

$$\overline{IMF_k}(t) = \frac{1}{n}\sum_{i=1}^{n} EMD_1\left(r_{k-1}(t) + \varepsilon_{k-1}EMD_{k-1}\left(\omega^i(t)\right)\right), \quad k = 2,3, \ldots K \tag{6}$$

$$r_k(t) = r_{k-1}(t) - \overline{IMF_k}(t) \tag{7}$$

（5）当残差不超过两个极值点且无法继续分解时，CEEMDAN 算法终止。总共得到 k 个 IMFs。最终的残差为 $R(t)$，原始信号 $f(t)$ 具有式（8）所示的关系。

$$f(t) = \sum_{k=1}^{k} \overline{IMF_k}(t) + R(t) \tag{8}$$

1.2 SE

SE[15] 是评估时间序列数据复杂性的方法，对于序列的 SE 而言，其熵值越大，序列越复杂。在下面的公式中，令 n 为数据点总数，m 为待比较序列的长度，然后可以形成向量 $x_m(i)$，如式（9）所示。

$$x_m(i) = [x(i), x(i+1), \ldots, x(i+m-1)], \quad i = 1,2, \ldots, n-m+1 \tag{9}$$

两个这样的向量之间的距离定义为 d_m：

$$d_m[x_m(i), x_m(j)] = \max\left[x_m(i+k) - x_m(j+k)\right], \quad 0 \le k \le m-1 \tag{10}$$

令 r 为接收矩阵的容差，v^m 为 $d_m[x_m(i), x_m(j)] \le r, i \ne j$ 的数量，ω^{m+1} 为 $d_m[x_{m+1}(i), x_{m+1}(j)] \le r, i \ne j$ 的数量，然后就可以确定匹配点的概率。$A^m(r)$ 是两个序列匹配 m+1 个点的概率，而 $B^m(r)$ 是匹配 m 个点的概率，如式（11）和（12）所示。

$$A^m(r) = \frac{1}{n-m}\sum_{i=1}^{n-m}\frac{1}{n-m+1}\omega^{m+1}(i) \tag{11}$$

$$B^m(r) = \frac{1}{n-m}\sum_{i=1}^{n-m}\frac{1}{n-m+1}v^m(i) \tag{12}$$

最后，样本熵定义为 $SampEn(m,r)$：

$$SampEn(m,r) = \lim_{n\to\infty}\left(-\ln\frac{A^m(r)}{B^m(r)}\right) \tag{13}$$

当 n 为有限值时，样本熵可以使用以下方程来估计：

$$SampEn(m,r,n) = -\ln\frac{A^m(r)}{B^m(r)} \tag{14}$$

1.3 PSO-VMD

VMD[16] 是一种被广泛使用的信号分解方法，其具体公式如下：

$$\min_{\{u_k\},\{\omega_k\}}\left\{\sum_k \left\| \partial_t\left[\left(\delta(t) + \frac{j}{\pi t}\right) * u_k(t)\right] * e^{-j\omega_k t}\right\|_2^2\right\}, \quad s.t.\sum_{k=1}^{K} u_k(t) = f(t) \tag{15}$$

式中：$f(t)$ 为原始数据，u_k 为有限带宽，w_k（k=2,3,⋯,K）为中心频率，K 为要分解的 IMF 的数量，$\delta(t)$ 为单位脉冲函数，t 为时间指标，∂_t 为 t 的偏导数，* 为卷积运算符。

为了解决优化问题，需要引入拉格朗日乘子 λ 和正则化

参数 α，然后将有约束变分问题转换为无约束变分问题，得到增广拉格朗日表达式：

$$L(\{u_k\},\{w_k\},\lambda)=\alpha\sum_k\left\|\partial_t\left[\left(\delta(t)+\frac{j}{\pi t}\right)*u_k(t)\right]e^{-jw_kt}\right\|_2^2+\left\|f(t)-\sum_k u_k(t)\right\|_2^2+\left\langle\lambda\left(t),f(t)-\sum_k u_k(t)\right)\right\rangle \tag{16}$$

通过交替方向乘子策略，可以计算出鞍点，得到 μ_k、w_k、λ 的解。原始 $f(t)$ 被分解为 K 个 IMFs，τ 是噪声容限，$\hat{u}_k^{n+1}(w)$、$\hat{u}_i(w)$、$\hat{f}(w)$、$\hat{\lambda}(w)$分别是 $u_k^{n+1}(t)$、$u_i(t)$、$f(t)$、$\lambda(t)$ 的傅里叶变换：

$$\begin{cases}\hat{u}_k^{n+1}(w)=\dfrac{\hat{f}(w)-\sum_{i\neq k}\hat{u}_i(w)+\widehat{\lambda}(w)/2}{1+2\alpha\left(w-\omega_k\right)^2}\\ \hat{u}_k^{n+1}=\dfrac{\int_0^\infty w|\hat{u}_k^{n+1}(w)|^2dw}{\int_0^\infty|\hat{u}_k^{n+1}(w)|^2dw}\\ \hat{\lambda}^{n+1}(w)=\hat{\lambda}^n(w)+\tau(\hat{f}(w)-\sum_k\hat{u}_k^{n+1}(w))\end{cases} \tag{17}$$

PSO[17] 是一种典型的群智能优化算法，汲取了自然界鸟群社交行为的启示，以便在解空间中探寻并识别出全局最优解。对于第 t 次迭代的第 i 个粒子，其位置和速度可以用公式写为：

$$V_{id}(t+1)=\omega V_{id}+c_1r_1\times\left(P_{id}-X_{id}(t)\right)+c_2r_2\times(P_{id}-X_{id}(t))$$
$$X_{id}(t+1)=X_{id}(t)+V_{id}(t)\ 1\le i\le n\ 1\le d\le N \tag{18}$$

式中：$V_{id}(t)$ 为第 t 次迭代后粒子 i 速度向量的第 d 维分量；$X_{id}(t)$ 为第 t 次迭代后粒子 i 位置向量的第 d 维分量；P_{id} 为粒子 i 历史最优解对应位置的第 d 维分量；P_{gd} 为群体最优解对应位置的第 d 维分量；c_1、c_2 为学习因子；r_1、r_2 为 [0,1] 之间的随机数。

VMD 的效果受模态分解的个数 (K) 和正则化参数 (α) 的影响很大，在进行变分模态分解时需要提前设置。在此之前，往往观察 K 取不同值时，各 IMFs 分量中心频率的变化，如果出现相似的分量则取最后的 K，因此以往的方法受人为等主观因素影响大，并且多次实验选取最佳的 VMD 参数往往会耗费巨大的时间成本，并不适合实际操作。在本研究中，使用了 PSO 与 VMD 结合的方法，旨在为股票价格数据的有效分解确定最优的 VMD 参数组合。PSO 算法的核心思想是确定一个适应度函数 [18]，以此计算粒子位置更新时对应的适应度值，通过比较新旧粒子的适应度值进行更新。由此可见，熵可以很好地反映复杂时间序列的动态特征，排列熵（PE）[19] 作为一种反映时间序列动态特性的有效工具，对于噪声具有较高的敏感性。高 PE 值表示较大的噪声水平，而较低的 PE 值则预示着更优的分解效果。本研究使用 PE 作为适应度函数，旨在评估由 VMD 分解所得的 IMFs 的复杂度及噪声水平。此外，IMFs 的 PE 平均值（MPE）被用作衡量每个粒子（即不同的 K 与 α 组合）性能的指标。具体算法流程如下。

（1）参数初始化。首先，设定 PSO 算法的基本参数，包括粒子群规模、最大迭代次数和学习因子。在此基础上，将 VMD 的两个关键参数 K 和 α 作为粒子的位置属性，从而初始化粒子群。

（2）初始种群的 MPE 计算。对于粒子群中的每一个粒子，利用其代表的 K 和 α 执行 VMD 分解，得到 K 个 IMFs。随后，计算这些 IMFs 的 PE，并据此求得 MPE。在种群中，所有粒子的 MPE 的最小值被确定为该种群的适应度评价标准。

（3）粒子位置和速度的更新。根据 PSO 算法的核心机制，更新每个粒子的位置和速度。

（4）MPE 的再计算与适应度更新。在粒子位置更新后，对每个粒子再次执行 VMD 分解，计算新的 MPE 值，并将其与先前的适应度值进行比较。若发现更低的 MPE 值，则相应地更新全局最优适应度值。

（5）迭代循环。持续执行步骤（3）和（4），直至达到预定的迭代次数上限。在每次迭代过程中，记录并更新当前迭代所得的最优适应度值及相应的粒子位置，即 K 和 α 值的最佳组合。

（6）输出最优解。在迭代过程完成后，输出达到最优适应度值时的粒子位置，这代表了 VMD 参数 K 和 α 的最优组合。

1.4 LSTM

LSTM[20] 在循环神经网络（RNN）的基础上精心设计并添加了记忆特性，以避免长依赖问题。LSTM 网络在隐藏层添加了遗忘单元和记忆单元，当新信息输入时丢弃次要信息，将重要信息保留在长期记忆中。LSTM 具有两种传输状态，分别是细胞状态 c_t 和隐藏状态 h_t。此外，为了解决梯度消失所带来的问题，LSTM 建立了门控机制来控制信息流，门控机制通常包括一个点乘法累加和一个神经网络层，其中 Sigmoid 激活函数 $\sigma(x)=\frac{1}{(1+e^{-x})}$ 或函数 $\tanh(x)=\frac{(e^x-e^{-x})}{(e^x+e^{-x})}$ 的神经网络层，用于确定可以传递多少信息。公式（19）～（24）为 LSTM 的详细计算公式，其中小写变量表示向量，大写变量表示矩阵。

$$f_t=\sigma\left(W_fx_t+U_f\hbar_{t-1}+b_f\right) \tag{19}$$
$$i_t=\sigma\left(W_ix_i+U_i\hbar_{t-1}+b_i\right) \tag{20}$$
$$\tilde{c_t}=\tanh\left(W_cx_t+U_c\hbar_{t-1}+b_c\right) \tag{21}$$
$$c_t=f_toc_{t-1}+i_to\tilde{c_t} \tag{22}$$
$$o_t=\sigma\left(W_ox_t+U_o\hbar_{t-1}+b_o\right) \tag{23}$$
$$\hbar_t=o_totanh(c_t) \tag{24}$$

LSTM 的输入向量为 $x_t\in\mathbf{R}^m$，隐藏状态向量为 $\hbar_t\in(-1,1)^n$，遗忘门的激活向量为$f_t\in(0,1)^n$，输入门的激活向量为$i_t\in(0,1)^n$，单元输入激活向量为$\tilde{c_t}=(-1,1)^n$。$c_t\in\mathbf{R}^n$ 为细胞状态向量，输出门的激活向量为$o_t\in(0,1)^n$，其中上标 m 和 n 分别指输入特征和隐藏单元的数量。对于$W_k\in\mathbf{R}^{n\times m}$，$U_k\in\mathbf{R}^{n\times n}$，$b_k\in\mathbf{R}^n$，$k=f,i,c,o$，分别表示输入向量 x_t 的权重矩阵、隐藏状态向量 h_t 的权重矩阵以及不同门或细胞状态 f、i、c、o 的偏置向量参数。此外，还使用特殊的矢量符号来更

好地解释，例如单元状态$c_t \in \mathbf{R}^n$不仅仅包含 LSTM 神经网络中的一个单元信息，而是包含 n 个 LSTM 单元的信息。

在实际计算中，使用适应性矩估计（Adam）[21] 优化算法，Adam 可以自适应地调整学习率且实现简单，其中参数的更新不受梯度变换的影响，因此它适用于噪声很大的股票价格数据集。此外，训练神经网络时经常遇到过拟合，因此需要设置 Dropout[22] 机制。Dropout 可以通过忽略特征检测器来减少神经元之间复杂的共适应关系，使网络学习更鲁棒的特征，从而解决过拟合问题。

2 数据

数据选取了 2017 年 6 月 27 日至 2024 年 1 月 19 日的 1600 个交易日数据，这些数据来源于同花顺官网的公开信息。将 2017 年 6 月 27 日至 2023 年 8 月 23 日的 1500 个交易日数据作为训练集，将 2023 年 8 月 24 日至 2024 年 1 月 19 日的 100 个交易日数据作为测试集。中国太保收盘价原始数据如图 1 所示，横轴表示日期，纵轴表示价格。需要注意的是，本研究假设每日的股票收盘价是连续的，并且与最近 30 天的股价紧密相关。

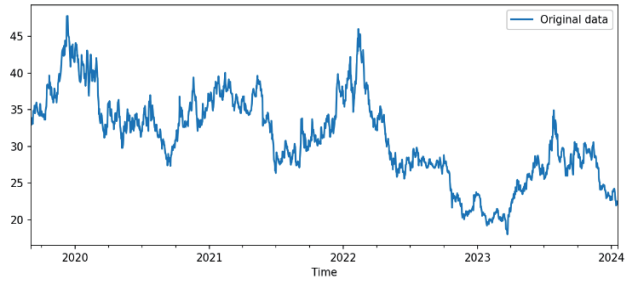


图 1 中国太保收盘价原始数据

本研究对股票收盘价的时间序列数据进行了综合统计分析，通过 Python3.11.5 和 Statsmodels 模块中的函数，使用统计检验来检验该股票数据的平稳性、自相关性和正态性。首先，对数据进行平稳性检验，通过增强型迪基 - 富勒（ADF）[23] 检验的应用，获得了一个 p 值，为 0.216 8，这个结果不足以拒绝非平稳的原假设。ADF 检验值为 -2.171 4，表明在 5% 的显著性水平上，数据未能证明其平稳性。其次，通过 Ljung-Box[24] 来评估该事件序列数据的自相关性，检验结果在 10 个滞后期内均显示 p 值显著小于 0.05，这表明收盘价数据存在自相关性。再次，为评估数据分布的正态性，采用 Jarque-Bera[25] 检验进行正态性检验，p 值为 0.000 01 小于 0.05，偏度为 0.020 9，峰度为 2.480 9，拒绝原假设，数据不服从正态分布。最后，通过 Statsmodels 模块计算得出的自相关函数（ACF）和偏自相关函数（PACF），如图 2 所示，提供了股票收盘价数据自相关性的直观证据。从 ACF 图可以观察到，存在逐渐减小的拖尾现象，这表明收盘价具有长期的自相关性。而 PACF 图表则揭示了在初期滞后期具有显著相关性后，迅速降至 0 的模式，尤其是在第二阶之后，数据的偏自相关性被截断，呈现出较为清晰的断层。

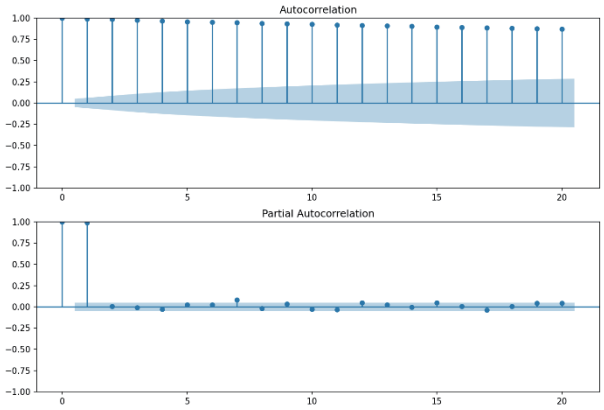


图 2 原始数据的 ACF 和 PACF

3 研究方法

对数据的综合统计检验结果表明，股票价格数据非平稳，噪声多，自相关性强，如果没有进行高效的降噪处理，就很难进行准确的预测。因此，本研究引入 CEEMDAN 和 PSOVMD，对其进行自适应分解和去噪，同时借鉴了 Wang 等人 [26] 的研究成果，引入样本熵整合来减少计算量，引入混合 LSTM 来稳定有效地预测股票价格，并将本研究所提出的模型称为 CEEMDAN-SE-PSOVMD- 混合 LSTM。具体而言，模型首先利用 CEEMDAN 将输入的原始数据分解为多个 IMFs；然后通过样本熵整合，将 IMFs 整合为高复杂度组和低复杂度组；对于高复杂度组，首先使用 PSOVMD 进行再分解，其次对 VMD 的分解结果使用集成 LSTM 进行预测，对于低复杂度组，使用各自 LSTM 进行预测；最后通过集成 LSTM 将预测结果进行拟合，得到最终的预测结果，模型如图 3 所示。

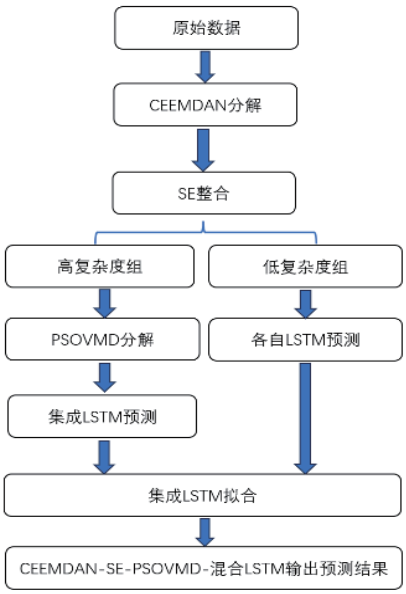


图 3 CEEMDAN-SE-PSOVMD- 混合 LSTM 模型预测模型

3.1 CEEMDAN 和样本熵整合

图 4 自上而下展示了原始数据，以及分解得到的 7 个本征模态函数（IMF0 至 IMF6）和最终的残差部分（即 IMF7）。这里，IMF7 被定义为残差。图 4 中横轴记录了股票价格的时间序列，而纵轴表示了不同部分的数值，单位是元。从 IMF0 至 IMF6，再到残差（IMF7），从上到下，可以清晰地看到时间序列的复杂度逐步降低，同时整体的价格趋势变得更加明确。

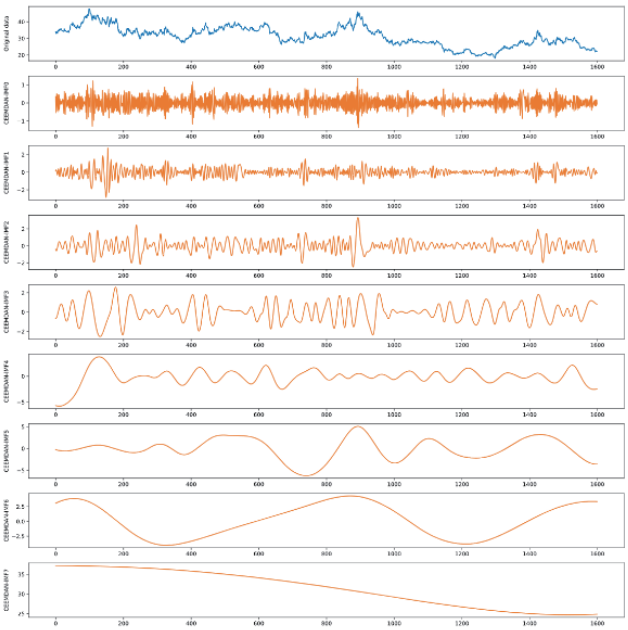


图 4 CEEMDAN 分解结果

为了提高模型的计算效率，将样本熵整合技术应用于 CEEMDAN 分解的结果。通过计算每个 IMF 的样本熵，能够直观地从图 5 中观察到 IMF0 和 IMF1 的样本熵值远高于其他 IMF。

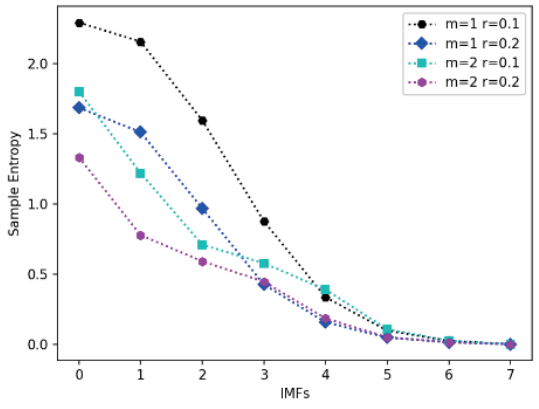


图 5 IMFs 样本熵值

相比之下，序列末尾 IMF5、IMF6 和 IMF7 样本熵值较低。基于样本熵揭示的复杂度的相似性，将这 8 个 IMF（包括残差）重新整合成 3 个新的本征模态函数群体，即：高频序列 Co-IMF0，包含 IMF0 和 IMF1；低频序列 Co-IMF1，包含 IMF2、IMF3、IMF4；趋势序列 Co-IMF2，包含 IMF5、IMF6、IMF7。本研究中将 Co-IMF0 称为高复杂度组，将 Co-IMF1 和 Co-IMF2 称为低复杂度组，得到的新的本征模态函数如图 6 所示。

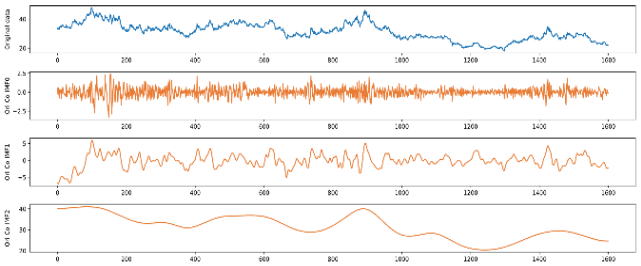


图 6 样本熵整合后的 Co-IMF

3.2 优化 VMD 和混合 LSTM 预测

原始数据在经过 CEEMDAN 分解和样本熵整合后，得到的第一个高频 Co-IMF0 包含大量噪声。高频噪声往往是股票价格预测的重中之重，K 和 α 是决定 VMD 效果的两个核心参数，K 值决定了 VMD 过程中将时间序列分解为多少个 IMFs。如果 K 值过大，可能会导致异常分解。将噪声或不相关的波动作为独立的模态分离出来，这样会降低分析的准确性和可解释性。如果 K 值过小，分解则无法捕获所有重要的频率组件，导致重要信息的丢失。α 控制了 VMD 算法中的平滑程度，α 值较大时，会增加分解的平滑程度，从而抑制一些短期波动，使 IMFs 更加集中于主要趋势；α 值较小时，分解得到的 IMFs 会包含更多高频波动，这意味着一些重要的趋势或周期性模式未被识别。因此，本研究提出使用 PSO 针对模型和股票价格数据的复杂度来寻找 VMD 的 K 和 α。迭代过程如图 7 所示，PSO 算法总共迭代 10 次。

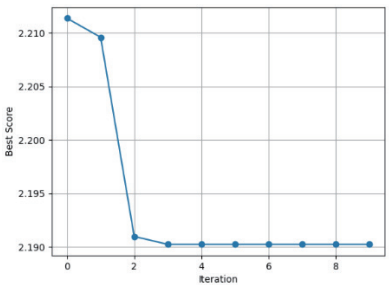


图 7 PSO 的迭代过程

在迭代 3 次后，迭代过程已经趋于稳定，10 次迭代结束后输出最优的 K 和 α，然后 VMD 对经过样本熵整合后的 Co-IMF0 进行进一步分解，并使用集成框架来预测分解出的结果，而对其他 Co-IMF(1,2) 则直接使用各自的 LSTM 预测方法。最后，将 Co-IMF 的所有预测结果再次通过集成 LSTM 方法进行拟合，以提高预测器的准确性和稳定性。PSO 算法的参数设置如表 1 所示，其中 G 为迭代次数，N 为 PSO 算法中的粒子数量，c_1 为认知因子，c_2 为社会因子，ω 为惯性权重，K 为模态分解的个数，α 为正则化参数。

表 1 PSO 参数设置

变量	G	N	c_1	c_2	ω	K	α
参数	10	10	1.5	1.5	0.5	[5-12]	[50-3000]

在 PSO 迭代结束后，K 值的输出为 9，即 PSO 算法寻找到的最优的 K 值为 9，所以 VMD 会分解出 9 个 IMFs（IMF0 ～ 8），VMD 分解结果如图 8 所示，最上面为输入 VMD 的高频序列（Co-IMF），下面依次为分解的 9 个 IMFs。因此，集成 LSTM 的参数的输入设置为 (None,30,9)，而对于各自的 LSTM 来说，输入则设置为 (None,30,1)，在混合模型最后一步，使用集成的 LSTM2 拟合前两个 LSTM 结果时，输入为 (None,31,3)。

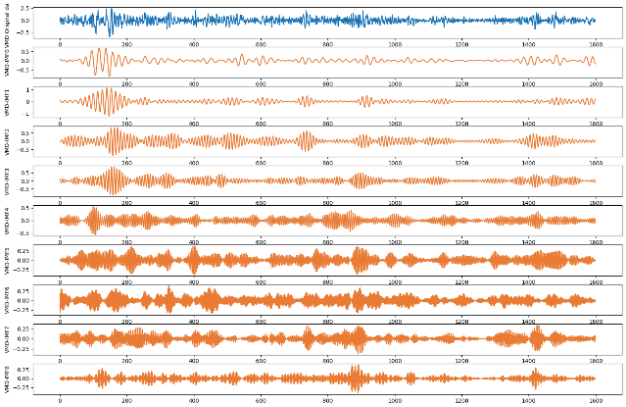


图 8 VMD 分解结果

3.3 股价预测

关于本研究所使用的研究方法具体步骤如下。

第一步：将原始数据输入混合模型，通过 CEEMDAN 分解为多个 IMFs。

第二步：通过 SE 对第一步分解得到 IMF 进行计算，并将具有相似熵值的 IMF 整合，得到 Co-IMF，分为高复杂度组（Co-IMF0）和低复杂度组（Co-IMF1,2）。

第三步：通过 PSO 优化后的 VMD，将第二步整合得到的具有高复杂度的（Co-IMF0）进行本文所称的再分解，然后使用集成 LSTM 对 VMD 分解得到的 K 个 IMF 进行预测，得到 Co-IMF0 的预测结果。然后，通过各自 LSTM 预测低复杂度的（Co-IMF1,2），得到低频序列的预测结果。

第四步：将所有预测结果再次通过集成 LSTM 方法进行拟合，以提高预测器的稳定性和准确性。

3.4 模型评估

在本研究中，设置了 4 个模型评价指标，分别是判定系数 (R^2)、均方根误差（RMSE）、平均绝对误差（MAE），以及平均百分比误差（MAPE），公式为：

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{25}$$

$$RMSE(y,\hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{26}$$

$$MAE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \tag{27}$$

$$MAPE(y,\hat{y}) = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i-\hat{y}_i|}{\max(\epsilon,|y_i|)} \tag{28}$$

R^2 越接近 1，表示模型的精度越高，当 R^2 为负数或者接近 0 时，意味着结果和数据集之间几乎没有相关性。RMSE 和 MAE 表示原始值和预测值之间的误差，RMSE 和 MAE 的值越接近 0，说明该模型的误差越小。MAPE 是与 RMSE 和 MAE 类似的百分比版本，MAPE 为 0 时表示模型完美，当输入数据有一些负数时，MAPE 无法正确计算结果，所以本文中的 MAPE 是在非标准化后计算的，以确保所有输入均为正值，以避免预测器可能会出现负值。

3.5 归一化

在研究中，为了提高预测器的计算速度和消除数据维度的影响，采用 min-max 来对数据进行归一化处理，公式为：

$$x = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{29}$$

式中：x 为原始数据，$x_{\max}$ 和 $x_{\min}$ 为数据集中特征向量的最大值和最小值。

4 结果

4.1 CEEMDAN-SE-PSOVMD- 混合 LSTM 模型预测结果

在输出结果方面，为了简化模型比较过程中的表述，本研究提出的 CEEMDAN-SE-PSOVMD 混合 LSTM 模型在各表和讨论中将统一被称为“本模型”。本模型预测了中国太保 2023 年 8 月 23 日至 2024 年 1 月 19 日为期 100 天的股票收盘价格，并绘制了预测值和真实值之间的拟合曲线，如图 9 所示。从图 9 中可以看出，本模型能够很好地捕捉原始数据中的内在规律和模式，与原始数据高度一致。

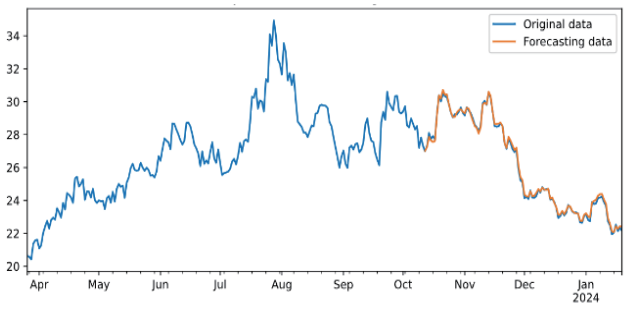


图 9 CEEMDAN-SE-PSOVMD- 混合 LSTM 拟合曲线

不仅如此，为了验证本模型在股价预测方面的性能，将其与当前主流的模型进行了比较，具体包括 CEEMDAN- 各自 LSTM、单一的卷积神经网络（single CNN）、单一的循环神经网络（single RNN）、单一的长短期记忆网络（Single

LSTM）和单一的支持向量回归（single SVR）。为了确保比较的准确性，每个模型均执行了 5 次独立运行，然后取这些运行的评估指标的平均值进行比较。比较结果在表 2 中展示，从中可以明显看出，本模型在预测效果上优于这些主流模型。

表 2 中国太保各模型评估指标对比

模型	R^2	MAE	RMSE	MAPE
本模型	0.997	0.111	0.135	0.431
CEEMDAN- 各自 LSTM	0.977	0.328	0.416	1.247
Single LSTM	0.970	0.345	0.478	1.314
Single CNN	0.958	0.348	0.505	1.487
Single RNN	0.967	0.367	0.484	1.405
Single SVR	0.765	1.212	1.342	4.826

此外，本研究还绘制了各模型预测结果与原始数据集的拟合曲线，如图 10 所示，进一步验证了提出模型的优越性。通过比较分析可以发现，与对比模型比较，本模型在拟合原始数据方面表现更佳，其余模型存在不同程度的偏差和滞后。通过对比发现，经过 CEEMDAN 噪声处理后的 LSTM 预测准确性要比单一的 LSTM 预测得更加准确，而单一的统计模型 Single SVR 与原始数据的拟合程度最为偏离。

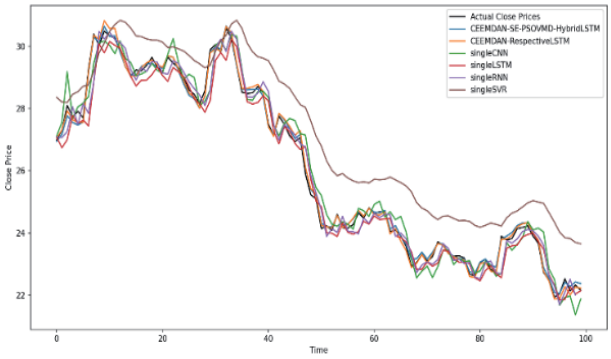
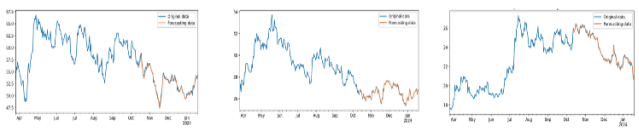


图 10 中国太保各模型拟合曲线

4.2 跨数据集验证的综合分析

为了全面评估本模型在股价预测方面的性能和鲁棒性，本研究不仅分析了中国太保的数据，还使用了春秋航空、山东黄金和伊利股份等三个不同行业的股票数据。选择 2017 年 6 月 27 日至 2023 年 8 月 23 日共 1500 个交易日的数据作为训练集，以及 2023 年 8 月 24 日至 2024 年 1 月 19 日的 100 个交易日数据作为测试集，假定每个交易日之间是连续的。

跨数据集的预测效果如图 11 所示，本模型与原始数据具有紧密的数据匹配。



春秋航空　伊利股份　山东黄金

图 11 跨数据集本模型拟合曲线

各项评估指标的对比结果见表 3，展示了本模型在不同数据集上相较于其他对比模型具有高准确率。通过在这些多样化的数据集上应用本模型，获得的预测结果凸显了本模型在跨行业股价预测中的性能，从而进一步证实了本模型的准确性和鲁棒性。

表 3 跨数据集的评估指标对比

数据	模型	R^2	MAE	RMSE	MAPE
春秋航空	本模型	0.975	0.302	0.377	0.569
	CEEMDAN- 各自 LSTM	0.924	0.514	0.666	0.964
	Single LSTM	0.924	0.382	0.576	1.122
	Single RNN	0.921	0.389	0.592	1.140
	Single CNN	0.916	0.472	0.606	1.385
	Single SVR	0.769	0.761	1.006	1.385
伊利股份	本模型	0.990	0.043	0.054	0.163
	CEEMDAN- 各自 LSTM	0.840	0.176	0.226	0.663
	Single LSTM	0.972	0.472	0.597	1.938
	Single RNN	0.974	0.428	0.571	1.174
	Single CNN	0.855	1.072	1.358	4.761
	Single SVR	0.728	1.691	1.860	7.219
山东黄金	本模型	0.995	0.084	0.106	0.356
	CEEMDAN- 各自 LSTM	0.975	0.181	0.245	0.765
	Single LSTM	0.871	0.397	0.530	1.268
	Single RNN	0.888	0.379	0.491	1.207
	Single CNN	0.880	0.403	0.513	1.277
	Single SVR	0.254	1.141	1.279	3.690

5 总结

本研究首先通过相关工作部分，详细回顾了混合模型在股票价格预测中的优势，然后通过基础理论和数据分析提出本研究的混合模型，并通过详细的实验过程证明了优化 VMD 的再分解方法对股票预测中噪声处理的重要性，证明了所提出的 CEEMDAN-SE-PSOVMD- 混合 LSTM 在股票价格方面的预测能力。最后将提出的模型与当下的各种模型进行对比，验证了本研究所提出模型的股价预测能力，并且研究还选取了不同行业的股票数据进行预测，通过将预测曲线与原始数据的曲线拟合程度对比，显示出了本混合模型在股票预测方面的性能和鲁棒性。

参考文献：

[1] 韩山杰，谈世哲．基于 TensorFlow 进行股票预测的深度学习模型的设计与实现 [J]. 计算机应用与软件，2018, 35(6): 267-271+291.

[2]LU W, LI J, WANG J, et al.A CNN-BiLSTM-AM method for stock price prediction[J].Neural computing and applications, 2021, 33: 4741-4753.

[3] 綦方中，林少倩，俞婷婷．基于 PCA 和 IFOA-BP 神经网络的股价预测模型 [J]. 计算机应用与软件，2020,37(1):116-121+156.

[4]PRASAD V V, GUMPARTHI S, VENKATARAMANA L Y, et al.Prediction of stock prices using statistical and machine

learning models:a comparative analysis[J].The computer journal,2022,65(5):1338-1351.

[5]VAN G T, SUYKENS J A K, BAESTAENS D E, et al.Financial time series prediction using least squares support vector machines within the evidence framework[J].IEEE transactions on neural networks,2001,12(4):809-821.

[6]LEE M C.Using support vector machine with a hybrid feature selection method to the stock trend prediction[J].Expert systems with applications,2009,36(8):10896-10904.

[7]QIU J, WANG B, ZHOU C.Forecasting stock prices with long-short term memory neural network based on attention mechanism[J].PloS one,2020,15(1):1-15.

[8]JING N, WU Z, WANG H.A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction[J].Expert systems with applications, 2021, 178:115019.1-115019.12.

[9]YANG Y, YANG Y, XIAO J.A hybrid prediction method for stock price using LSTM and ensemble EMD[EB/OL].(2020-12-04)[2024-04-02].https://doi.org/10.1155/2020/6431712.

[10]PAI P, LIN C.A hybrid ARIMA and support vector machines model in stock price forecasting[J].Omega, 2005,33(6):497-505.

[11]RIBEIRO G T, SANTOS A A P, MARIANI V C, et al.Novel hybrid model based on echo state neural network applied to the prediction of stock price return volatility[J].Expert systems with applications,2021,184:115490.1-115490.13.

[12]ZHANG L, WANG J, NIU X, et al.Ensemble wind speed forecasting with multi-objective archimedes optimization algorithm and sub-model selection[J].Applied energy, 2021, 301: 117449.1-117449.18.

[13]ZHOU F, HUANG Z, ZHANG C.Carbon price forecasting based on CEEMDAN and LSTM[J].Applied energy,2022,311:118601.1-118601.20.

[14]TORRES M E, COLOMINAS M A, SCHLOTTHAUER G, et al.A complete ensemble empirical mode decomposition with adaptive noise[C]//2011 IEEE International Conference on Acoustics, Speech and Signal Processing,[v.1].Piscataway: IEEE, 2011:4144-4147.

[15]RICHMAN J S, MOORMAN J R.Physiological time-series analysis using approximate entropy and sample entropy[J]. American journal of physiology-heart and circulatory physiology, 2000, 278(6):H2039-H2049.

[16]DRAGOMIRETSKIY K, ZOSSO D.Variational mode decomposition[J].IEEE transactions on signal processing, 2013, 62(3):531-544.

[17]EBERHART R, KENNEDY J.Particle swarm optimization[C]//Proceedings of ICNN'95-International Conference on Neural Networks.Piscataway: IEEE, 1995: 1942-1948.

[18]ZHOU F, YANG X, SHEN J, et al.Fault diagnosis of hydraulic pumps using PSO-VMD and refined composite multiscale fluctuation dispersion entropy[J].Shock and vibration, 2020, 2020: 8840676.1-8840676.13.

[19]BANDT C, POMPE B.Permutation entropy:a natural complexity measure for time series[J].Physical review letters, 2002, 88(17):4102-4103.

[20]HOCHREITER S, SCHMIDHUBER J.Long short-term memory[J].Neural computation,1997,9(8):1735-1780.

[21]KINGMA D P, BA J.Adam:a method for stochastic optimization[EB/OL].(2014-12-22)[2024-04-05].https://doi.org/10.48550/arXiv.1412.6980.

[22]HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et al.Improving neural networks by preventing co-adaptation of feature detectors[EB/OL].(2012-07-03)[2024-04-10].https://doi.org/10.48550/arXiv.1207.0580.

[23]CHEUNG Y W, LAI K S.Lag order and critical values of the augmented dickey–fuller test[J].Journal of business & economic statistics,1995,13(3):277-280.

[24]LJUNG G M, BOX G E P.On a measure of lack of fit in time series models[J].Biometrika,1978,65(2):297-303.

[25]JARQUE C M, BERA A K.Efficient tests for normality, homoscedasticity and serial independence of regression residuals[J]. Economics letters,1980,6(3):255-259.

[26]WANG J, SUN X, CHENG Q, et al.An innovative random forest-based nonlinear ensemble paradigm of improved feature extraction and deep learning for carbon price forecasting[J]. Science of the total environment,2021,762:143099.

【作者简介】

王艺涵（2000—），男，山东淄博人，硕士研究生，研究方向：数据分析、人工智能。

路翀（1966—），通信作者（email：498841300@qq.com），男，江苏扬州人，博士，教授，硕士生导师，研究方向：人工智能、数据分析、图像处理。

龙洁（2001—），女，湖南衡阳人，硕士研究生，研究方向：数据分析、人工智能。

雷一鸣（2001—），女，河南洛阳人，硕士研究生，研究方向：数据分析、人工智能。

（收稿日期：2024-05-11）