Advances in Applied Mathematics 应用数学进展, 2023, 12(12), 5175-5185
Published Online December 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.1212508

Hans汉斯

基于PCA、ICEEMDAN和LSTM的股票价格预测混合框架

刘玉昆

长沙理工大学数学与统计学院，湖南 长沙

收稿日期：2023年11月25日；录用日期：2023年12月19日；发布日期：2023年12月26日

摘　要

投资者在市场上买卖股票的目的是为了获得最大的回报。然而，股票价格表现出非线性和非平稳性，难以准确预测。为了解决这个问题，结合主成分分析(PCA)，完全自适应噪声集合经验模态分解(ICEEMDAN)和长短期记忆网络(LSTM)，制定了一个混合预测模型，称为PCA-ICEEMDAN-LSTM，以预测中国股票指数收盘价。在这项研究中，选取8个股市中常用的技术指标作为原始特征，利用PCA筛选出最符合预期的几个技术指标作为LSTM的输入特征，ICEEMDAN分解得到的分量作为目标变量。对2018~2022年中国股票市场价格的实验进行了研究，并使用各种统计指标作为评估标准。实验得到的结果表明，该框架产生了最好的性能相比，基线方法预测股票市场价格。此外，采用PCA和ICEEMDAN可以帮助提高基线LSTM模型的性能。

关键词

主成分分析，经验模态分解，长短期记忆网络

A Hybrid Framework for Stock Price Prediction Based on PCA, ICEEMDAN, and LSTM

Yukun Liu

School of Mathematics and Statistics, Changsha University of Science & Technology, Changsha Hunan

Received: Nov. 25th, 2023; accepted: Dec. 19th, 2023; published: Dec. 26th, 2023

Abstract

The purpose of investors buying and selling stocks in the market is to achieve the maximum re-

文章引用: 刘玉昆. 基于 PCA、ICEEMDAN 和 LSTM 的股票价格预测混合框架[J]. 应用数学进展, 2023, 12(12): 5175-5185.
DOI: 10.12677/aam.2023.1212508

turn. However, stock prices exhibit non-linearity and non-stationarity, making accurate prediction challenging. To address this issue, a hybrid forecasting model, named PCA-ICEEMDAN-LSTM, was developed by combining Principal Component Analysis (PCA), Improved Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (ICEEMDAN), and Long Short-Term Memory networks (LSTM) to predict the closing prices of Chinese stock indices. In this study, eight commonly used technical indicators in the stock market were selected as the original features. PCA was utilized to filter out the most relevant technical indicators as input features for LSTM, and the components obtained from ICEEMDAN decomposition were used as target variables. Experiments were conducted on the Chinese stock market prices from 2018 to 2022, and various statistical indicators were used as evaluation criteria. The results obtained indicate that this framework produces the best performance compared to baseline methods in predicting stock market prices. Furthermore, the use of PCA and ICEEMDAN helps to enhance the performance of the original LSTM model.

Keywords

Principal Component Analysis (PCA), Improved Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (ICEEMDAN), Long Short-Term Memory (LSTM)

1. 介绍

长期以来，追求准确及时的股票价格预测一直是金融研究的前沿课题之一。由于股票市场固有的波动性，再加上众多经济、政治和公司特定因素的复杂相互作用，股票价格预测是一项极具挑战性的任务。金融监管机构和政策制定者正在共同努力，利用先进的预测工具来避免金融危机。股票价格预测技术经历了创新演变，从时间序列分析方法[1]过渡到机器学习[2]，并进一步发展到深度学习[3]。

对于金融数据作为一种特殊的时间序列，传统的计量经济学模型如 ARIMA 和 Goach 一直被用于价格预测[4]，但金融数据的非线性和非光滑性导致这类模型的预测有很多局限性[5]。近年来，随着机器学习技术的飞速发展，一些经典算法在股价预测中得到了广泛的应用。将支持向量机(SVM)与 KNN 相结合的融合框架应用于沪深股指的价格回归预测[6]。随机森林已被用于股票市场趋势预测，准确率为 85%~95% [7]。众多学者和专家对启发式算法在股票预测中的应用进行了探索。金等人在[8]开发了一个智能决策支持系统，以确定最佳的交易规则，然后采用制定最佳的买入或卖出策略。Das 等人[9]提出了一种基于进化框架的萤火虫算法，通过对 OSELM 模型的变换来最小化特征，从而提高未来股价预测的准确性。SrijiranonK 等人[10]提出了一种基于主成分分析(PCA)和长短期记忆(LSTM)的股票价格预测模型，实验论证该模型能够准确预测股价波动趋势。

本文的结构如下：第 2 章介绍研究所需要用到的理论知识；第 3 章介绍模型的构建思路和基本流程；第 4 章介绍实验论证所提出模型的优越性，并设置两个模型来进一步说明。

2. 背景理论

本部分介绍了本研究所使用的相关理论。包括主成分分析、完全经验模式分解和长短期记忆网络。

2.1. 主成分分析(PCA)

主成分分析(PCA)是最知名的减少消耗的技术之一[11]。PCA 是一种特征变换方法，用于通过将许多变

量转化为更少的变量来降低海量数据集的维度，同时保留大集合中的大部分信息。这种技术节省了运行模型的资源并提高了准确性。在股票预测领域，由于技术指标取决于趋势、波动率、成交量、动量和每日回报，因此它们可以推广到各种场景[12]。PCA 可以将大量技术指标视为输入特征，而不会遇到维数灾难。PCA 的优点可以应用于各种数据源和应用，例如游客行为分析和海上风力涡轮机选择。此外，一些研究表明，将机器学习和 PCA 相结合可以显著改善模型，特别是与成熟的降维技术相比。PCA 的基本步骤如下：

1) 对原始数据进行标准化，以确保每组数据对分析的贡献相等。在数学上，归一化方程表示为式(1)，其中 $x_{\min}$ 和 $x_{\max}$ 分别表示特征的最小值和最大值，x 表示原始值，而 $x_{\text{normailzed}}$ 表示归一化后的新值。

$$x_{\text{normailzed}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

2) 根据归一化数据矩阵建立协方差矩阵。由于数据集是 n 维的，这将导致一个表示为矩阵 A 的 $n \times n$ 维的协方差矩阵。

3) 第三步是计算协方差矩阵的特征向量和特征值来识别主成分。矩阵 A 的特征值 λ 通过求解式(2)来找到，其中 I 表示与 A 相同维的单位矩阵，目的是满足矩阵减法的基本要求。对于每个 λ，可以通过求解式(3)来找到对应的特征向量 v。

$$\det\left(\lambda I - A\right) = 0 \tag{2}$$

$$\left(\lambda I - A\right) v = 0 \tag{3}$$

4) 通过将具有相应特征值的特征向量从最大到最小排序来减少原始矩阵。具有最高特征值的特征向量成为数据的主成分。在此之后，首先选择 p 个特征值以降低维度，然后接收主成分。

2.2. 改进完全集合经验模态分解(ICEEMDAN)

ICEEMDAN 是一种时域信号分解方法，是 EMD 的改进版本。为了提高后者对模态混叠的敏感性，提出了 CEEMDAN 分解来消除噪声的影响，以减少模态混叠；为了进一步改进 CEEMDAN 分解，提出了 ICEEMDAN 分解，它改进了噪声添加策略，以自适应地添加噪声。ICEEMDAN 还包括对噪声添加的优化，以提高分解质量和效率。

2.3. 长短期记忆神经网络(LSTM)

LSTM 的门结构一共有 3 个，分别是遗忘门(Forget Gate)、更新门(Input Gate)以及输出门(Output Gate)[13]。相比于原始的 RNN 的隐含层(Hidden State)，LSTM 增加了一个细胞状态 C_t (Cell State)。神经元结构图如图 1 所示。它们对应的计算公式分别为：

$$\begin{aligned} F_t &= \sigma\left(W_f \cdot \left[H_{t-1}, X_t\right] + b_f\right) \\ I_t &= \sigma\left(W_i \cdot \left[H_{t-1}, X_t\right] + b_i\right) \\ O_t &= \sigma\left(W_o \cdot \left[H_{t-1}, X_t\right] + b_o\right) \end{aligned} \tag{4}$$

其中，H_{t-1} 是上一时刻的隐藏状态，σ 为激活函数，通常是 sigmoid 函数，F_t、I_t 和 O_t 分别为遗忘门、输入门和输出门状态结算结果，W_f、W_i 和 W_o 分别为遗忘门、输入门和输出门的权重矩阵，b_f、b_i 和 b_o 分别为遗忘门、输入门和输出门的偏置项，LSTM 的最终输出由输出门和单元状态共同确定。

在图 1 中，$\tilde{C}_t$ 为候选值向量，输入值和候选值向量的乘积用来更新细胞状态，计算过程如下：

$$\tilde{C}_t = \tanh\left(W_c H_{t-1} + W_c x_t + b_c\right) \tag{5}$$

$$C_t = F_t C_{t-1} + I\tilde{C}_t \tag{6}$$

$$H_t = O_t \tanh\left(C_t\right) \tag{7}$$

$$F\left(x\right) = \frac{1}{1+\mathrm{e}^{-x}} \tag{8}$$

$$F\left(x\right) = \tanh\left(x\right) \tag{9}$$

其中，W_c 为输入单元状态权重矩阵，b_c 为输入单元状态偏置项，tanh 为激活函数。遗忘门控制当前时刻细胞状态丢弃信息的多少，O_t 为神经元输出值，H_t 为当前时刻隐藏状态。

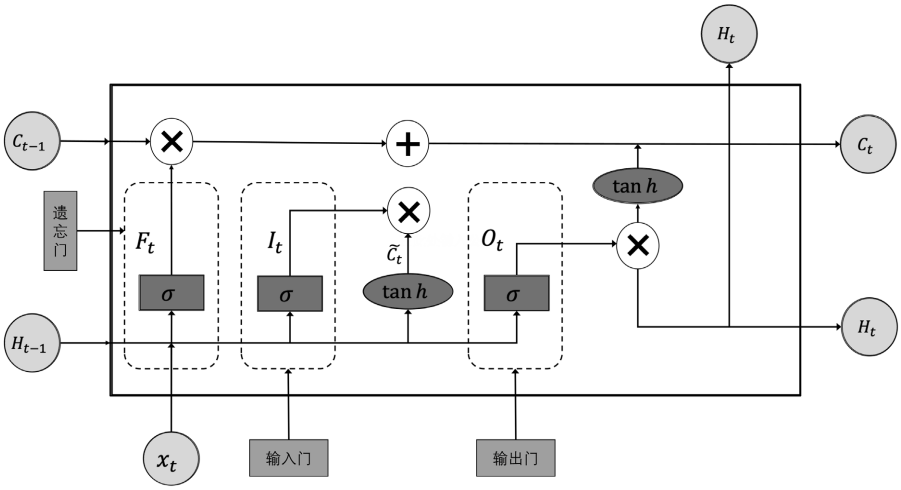

Figure 1. LSTM neuron structure diagram
图 1. LSTM 神经元结构图

3. 模型介绍

本研究的目的是使用 PCA、ICEEMDAN 和 LSTM 的组合为中国股票市场的收盘价提出一个混合框架。所提出的模型的总体架构如图 2 所示。该架构分为特征工程和预测模型两个部分。

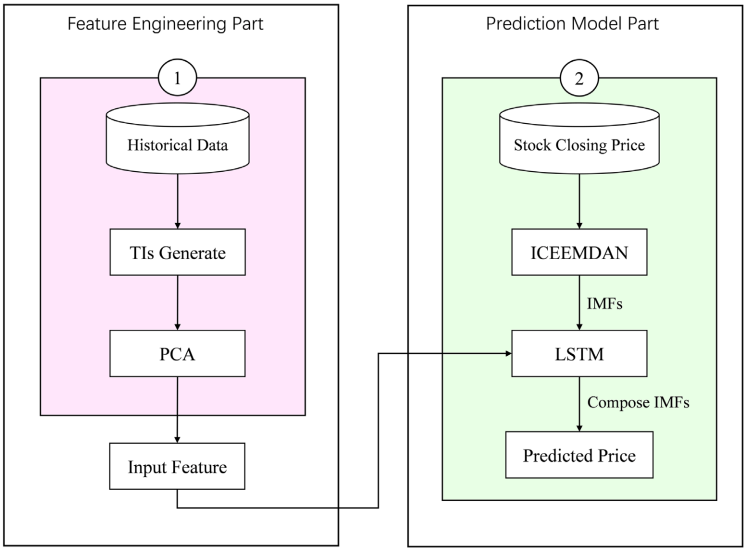

Figure 2. The specific process of the proposed fusion framework
图 2. 所提出融合框架的具体流程

3.1. 特征工程

本小节描述了为预测模型构建输入特征的过程。本研究选取了 8 个技术指标作为模型的输入特征。简单地说，特征越多，过度适应的风险就越高。为了解决这一问题，采用主成分分析对特征空间进行精简，同时考虑一组主特征。为了从 PCA 创建主成分，需要遵循图 3 所示的步骤。首先，从 https://www.tushare.pro/ 获得历史数据，包括开盘、高点、低点、收盘价和成交量数据。此后，使用 https://github.com/bukosabino/ta 中的“ta”包来生成技术指标。然后，在使用主成分分析进行数据降维之前，对技术指标进行归一化处理。主成分分析的结果是主成分，在本研究中，主成分是从第一个主成分开始，直到解释的方差比率之和大于 95%。因此，这表明技术指标中 95%的信息可以用“*n*-principl”原理来解释。

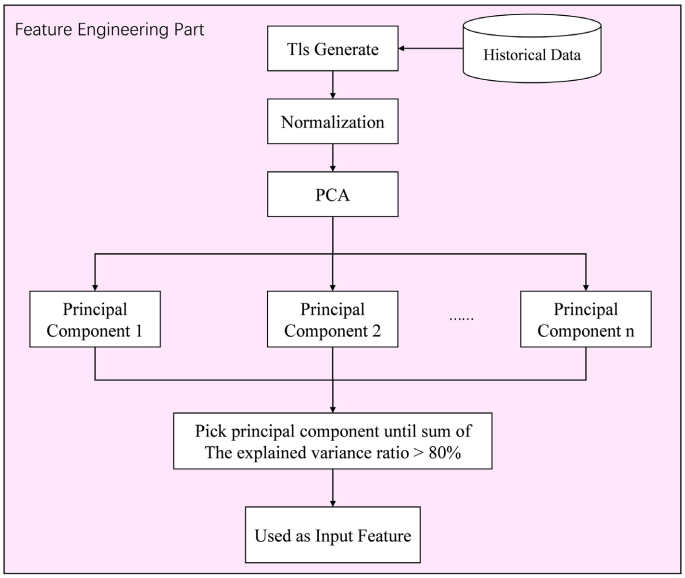

Figure 3. Constructing principal components of technical indicators
图 3. 技术指标主成分构建过程

3.2. 预测模型

本研究提出了一种基于 ICEEMDAN 和 LSTM 相结合的集成预测模型，以最大限度地提高预测效果，并最大限度地降低计算复杂度。所提出的模型如图 2 所示，包括以下四个步骤：

1) 首先，应用 ICEEMDAN 算法将原始股票收盘价时间序列分解为多个独立的 IMF 分量和一个残差分量 Residue。

2) 其次，将来自特征工程部分的主成分作为模型的输入特征。

3) 接着，LSTM 模型被用作每个 IMF 分量的预测工具，相应的分量获取相应的预测值。其中，LSTM 是由每个 IMF 单独训练的，因此，网络参数、神经元数目和批量大小等参数都是为每个 IMF 专门调整的。这是混合 ICEEMDAN-LSTM 模型优于单一 LSTM 模型的显著差异。

4) 最后，在获得 IMF 的预测结果后，使用组合每个预测 IMF 以获得最终的预测股票收盘价。然后，使用性能指标与其他模型的结果进行了比较。

3.3. 评价指标

为了论证所提出模型的好处，本研究采用了四个指标：平均绝对误差(MAE)、均方误差(MSE)、均方

根误差(RMSE)和决定系数(R^2)。令 y_i 表示实际数据，$\hat{y}_i$ 表示预测结果，其中 N 表示时间序列的长度。详细描述见表 1。

Table 1. Evaluation metrics
表 1. 评价指标

指标名称	定义	公式
MAE	平均绝对误差	$\frac{1}{N}\sum_{i=1}^{N}\left\|(y_i-\hat{y}_i)\right\|$
MSE	均方误差	$\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2$
RMSE	均方根误差	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}$
R^2	决定系数	$1-\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y}_i)^2}$

4. 实验部分

本节介绍利用所提出混合模型来预测中国股市股票指数价格的实验内容。

4.1. 数据收集

本研究以中国股市收盘价为例，以一步预测验证所提出模型的预测准确性。从 https://www.tushare.pro/ 获得包括开盘价、当日最高价、当日最低价、成交量、成交额、收盘价，时间跨度是从 2018 年 1 月 2 日到 2022 年 12 月 30 日，只选取交易日的数据用于研究。所选数据的收盘价走势在图 4 中可视化。

Figure 4. The daily closing price of CSI 300
图 4. 沪深 300 的每日收盘价

4.2. 技术指标

技术指标是计算机系统按照一定的数学统计方法，运用一定的数学计算公式或定量模型，生成的一定的指标值或图形曲线。用指数技术判断股价未来走势的分析方法，就是技术指数分析法，属于技术分析。本研究选取了与股市收盘价相关的 8 项技术指标。类别和名称见表 2。

Table 2. Technical indicators
表 2. 技术指标

类别	描述
趋势	区间震荡线(Detrended Price Oscillator, DPO) 简单移动平均线(Simple Moving Average, SMA)
交易量	负量指标(Negative Volume Index, NVI)
波动性	布林带宽度(Bollinger Bands Width, BOLL)
动量	随机相对强弱指标(The Stochastic RSI) 价格变动率(Price Rate of Change, PRC) 百分比成交量震荡指标(Percentage Volume Oscillator-Histogram, PVO) 威廉姆斯指数(Williams %R)

4.3. 实验结果

ICEEMDAN 分解分量及其预测结果

在建立预测模型时，利用 ICEEMDAN 将股票收盘价格作为历史数据转化为新的数据。如图 5 所示，

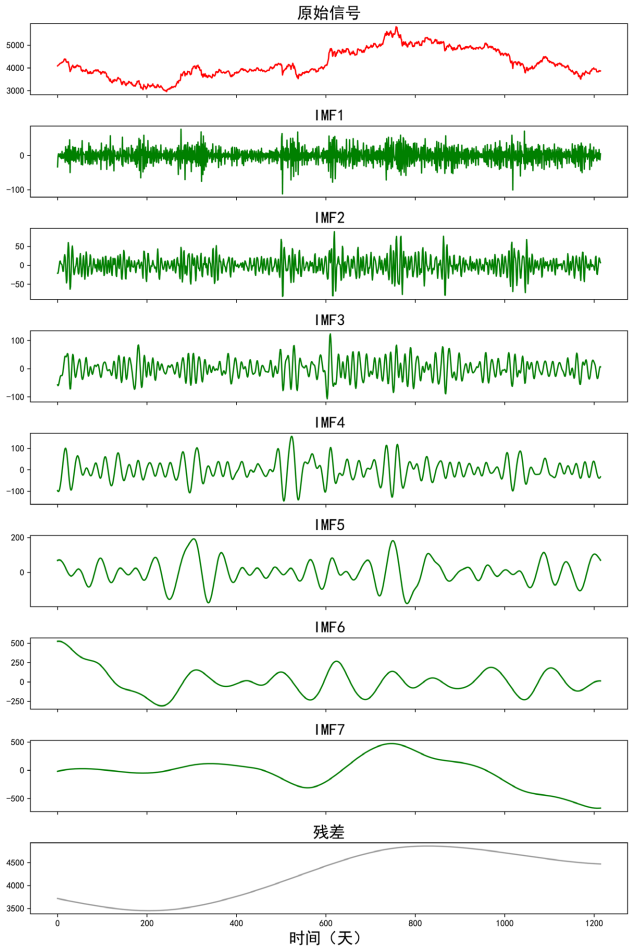

Figure 5. Original closing price and decomposed IMFs
图 5. 原始收盘价和分解后的 IMFs

演示了使用 ICEEMDAN 创建 IMF 的分解结果。其中的 6 个 IMF 是从最初的收盘价序列和从高到低频率的顺序来分解的。然而，根据原始数据，IMF 的数量是不同的。重复 ICEEMDAN 过程，直到在图 5 中的残差项上只有一个全局最大值和最小值。如果原始数据更改，IMF 的数量也将更改。另一方面，当 ICEEMDAN 应用于相同的数据时，IMF 的数量仍然是相同的值。另外还需要注意的是，IMF 是通过从原始收盘价中减去得到的，因此所有 IMF 的总和与原始的完全相同。因此，所有 IMF 的预测结果的总和可以被认为是对原收盘价的预测结果，本研究后续得到的最终结果基于该思路。

图 5 表明，它可分为三组。第一组是原始数据中的高频分量。这一组是由最初的几个 IMF 代表的，噪音很大。第二组是中频分量。由具有中等噪声的中心 IMF 表示。最后一组是低频分量。这一组由最近几个几乎没有噪音的 IMF 代表。而且，最后的残差项是和一只股票的走势相媲美的。假设 LSTM 能够准确地预测低频 IMF，但它不能很好地预测高频 IMF。为了最大限度地提高预测效率，LSTM 是由每个 IMF 单独训练。因此，每个 IMF 的超参数、隐含层数和权重也是不同的。

本研究选取的技术指标共有 8 个，它们对应的类共有 22 个，将它们都用作特征不仅会加大计算量，还容易导致过拟合，因此，我们采用了 PCA 对 22 个待选特征进行筛选，实验得到 7 个主成分，即他们的方差贡献率大于等于 95%，如图 6 所示。接着将对收盘价进行 ICEEMDAN 分解得到的 7 个 IMF 和 1 个残差项分别作为目标变量，单独训练每一个 LSTM，训练所用的超参数见表 3。

对每一个 LSTM 来说，我们统一设置数据集的 30%为训练集，70%为测试集，经过 PCA 降维后，得到的 7 个主成分作为输入特征，每一个 LSTM 中 IMFs 分别作为目标变量，训练后的预测结果如图 7 所示。从图 7 中可以看到，紫色虚线与红色实线拟合程度更好，其次是绿色虚线，最差的是红色虚线，也就是说，本文提出的 ICEEMDAN-PCA-LSTM 的预测效果优于 ICEEMDAN-LSTM 和 LSTM 模型。

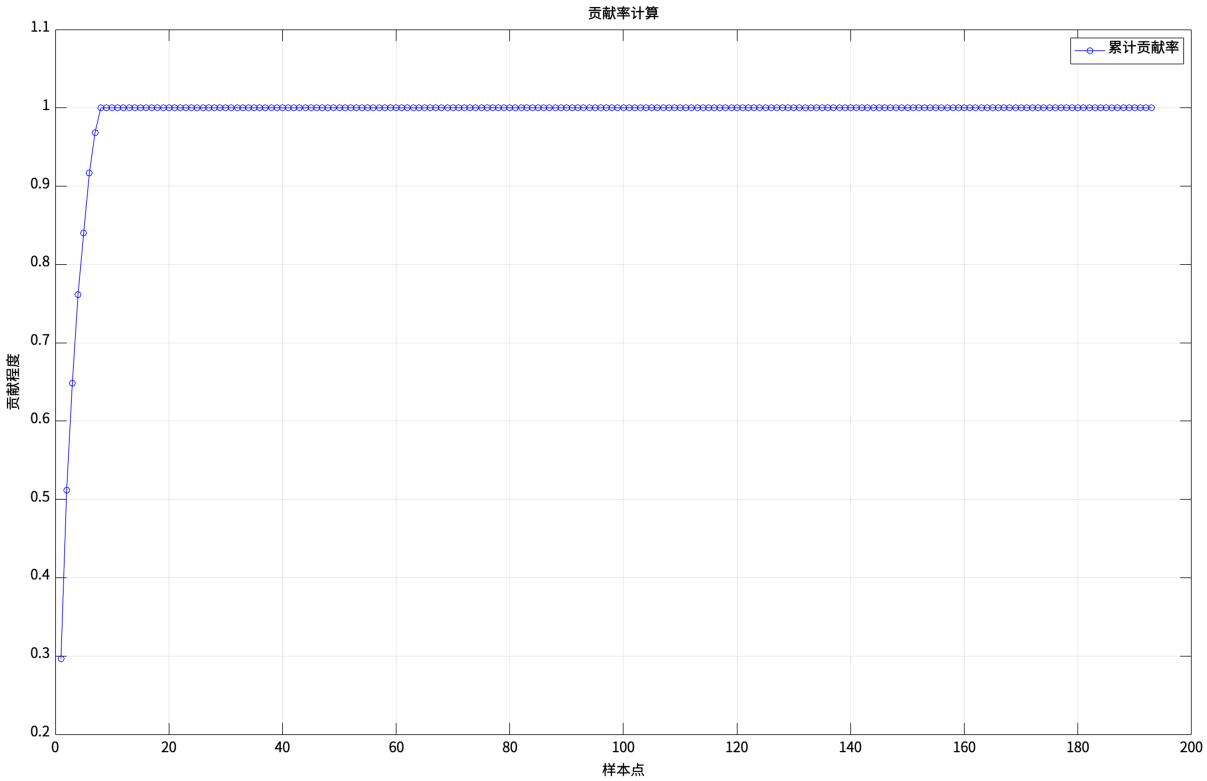

Figure 6. Change in principal component variance contribution rate
图 6. 主成分方差贡献率变化

Table 3. Configuration of hyperparameters for each LSTM
表 3. 各个 LSTM 的超参数配置

IMFs	最大训练次数	梯度阈值	初始学习率	学习率下降因子	学习率下降周期	正则化参数
1	150	1	0.01	0.02	400	0.001
2	150	1	0.01	0.02	400	0.001
3	150	1	0.02	0.2	400	0.0005
4	150	1	0.01	0.2	400	0.001
5	200	1	0.005	0.2	400	0.001
6	150	1	0.01	0.2	400	0.0005
7	150	0.4	0.01	0.001	600	0
Res	150	1	0.00095	0.4	200	0

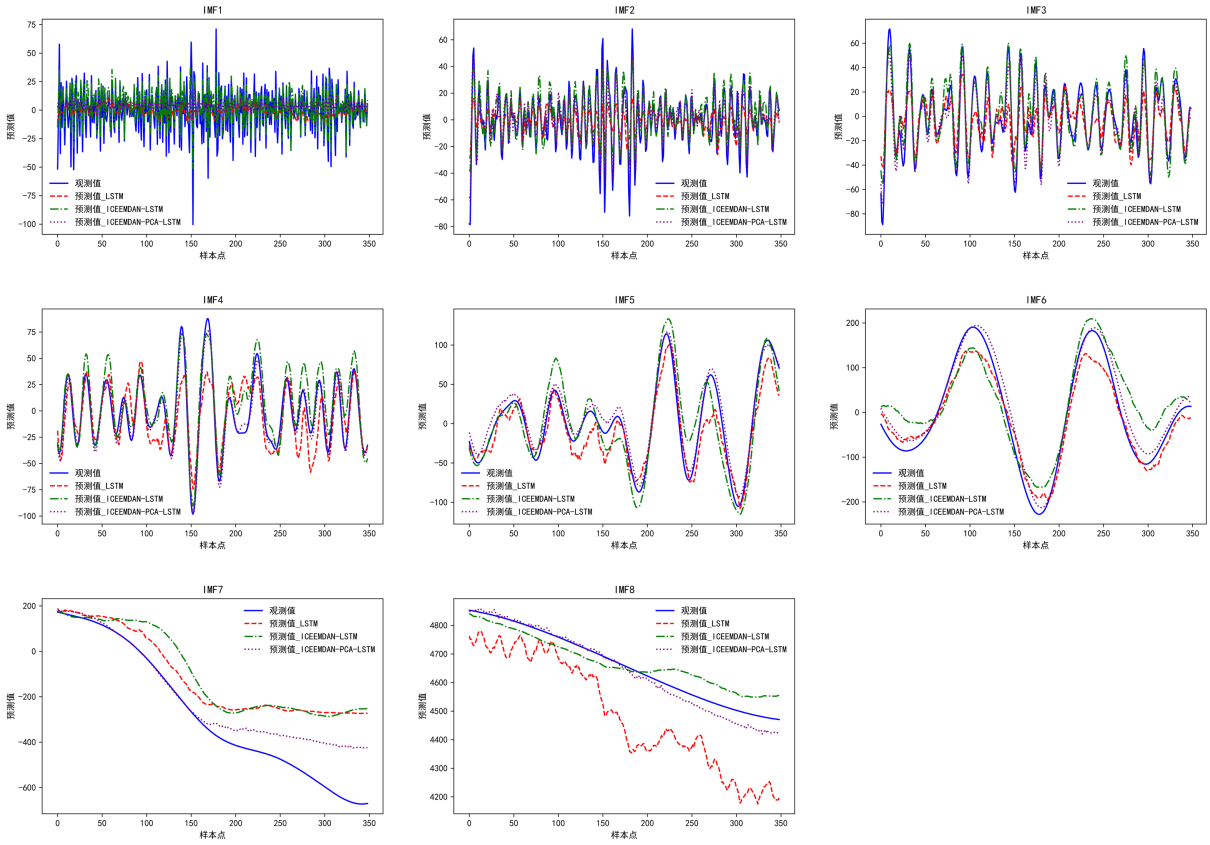

Figure 7. Comparison of IMFs prediction results figs
图 7. IMFs 预测结果对比

接下来，本研究将以上各个 IMF 分量的预测结果对应相加，最终得到原始数据集的预测结果对比图 8。与预期结果基本吻合，本文提出的 ICEEMDAN-PCA-LSTM 的预测效果最佳，其次是 ICEEMDAN-LSTM，最差的是 LSTM，得到它们的评价指标如表 4。

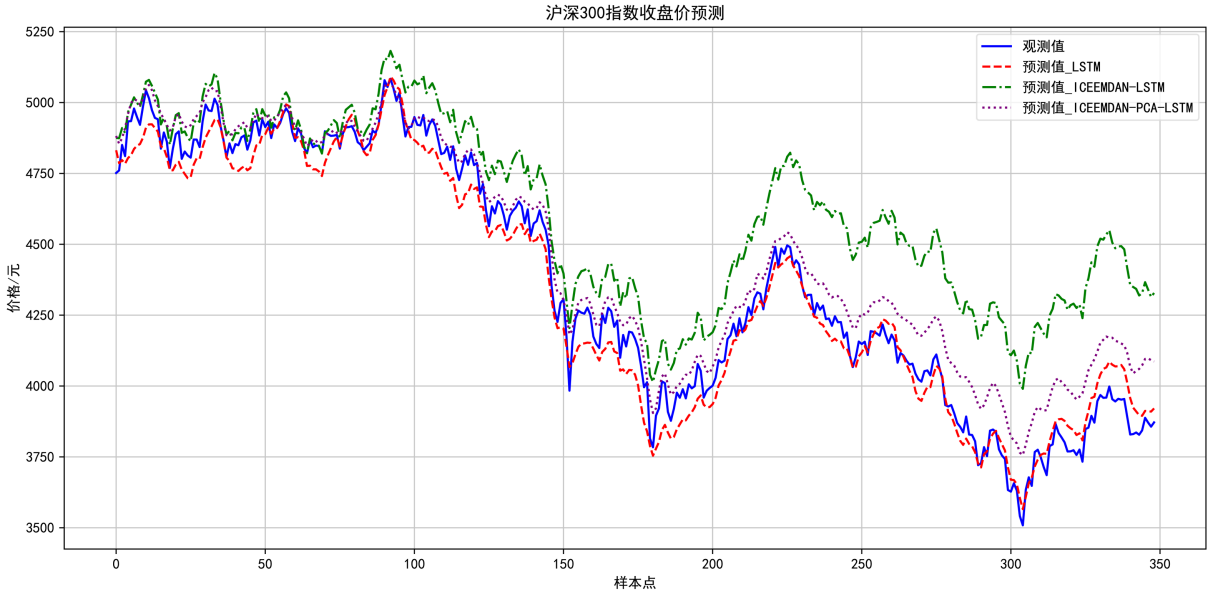

Figure 8. CSI 300closing price forecast
图 8. 沪深 300 指数收盘价预测

Table 4. Comparison of four evaluation indicators
表 4. 3 种评价指标对比

模型	MAE	MSE	RMSE	R^2
LSTM	237.2117	83925.1739	289.6984	0.5580
ICEEMDAN-LSTM	87.8491	12039.4106	109.7242	0.9366
ICEEMDAN-PCA-LSTM	**59.3076**	**4889.4412**	**69.9245**	**0.9743**

5. 结论与讨论

本研究旨在探讨股票价格时间序列的预测问题，选取了 Tushare 大数据社区提供的沪深 300 指数历史数据作为研究样本。运用数学统计方法计算了如 MACD、RSI、MA 等多种技术指标，这些指标基于历史的收盘价、开盘价、当日最高价和最低价等基础数据，以反映股票市场的波动特征。

研究初步选定了八个技术指标作为预测模型的输入变量，并通过 PCA 主成分分析方法进行降维处理，筛选出主成分贡献率超过 95%的技术指标，作为最终的输入特征。同时，采用完全自适应经验模态分解法(ICEEMDAN)对收盘价进行分解，得到七个本征模态函数(IMF)和一个残差项。以 PCA 筛选的主成分变量作为特征，七个 IMF 和残差项作为目标变量，分别构建八个长短期记忆(LSTM)网络进行训练和预测。预测结果通过重构相加，得出最终的收盘价预测值。

为了验证模型效果，本研究设立了两个对照模型，并选用四个评价指标作为模型效能的衡量。模型比较的图示展示于图 8，评价指标的比较结果列示于表 4。通过对比分析，可以看出三种模型的预测效果呈递增趋势。其中，ICEEMDAN 方法能有效提升 LSTM 预测的准确度，这是因为在未分解的收盘价时间序列中含有大量噪声，这些噪声若直接用 LSTM 预测会影响结果。ICEEMDAN 能有效压制噪声，保留时间序列中更多信息。此外，PCA 主成分分析在输入特征的降维过程中，排除了那些贡献率较低的特征，这些特征可能增加模型复杂度且不增加预测效果，因此最佳的策略是将它们排除。在构建数据集之前无法识别并排除这些特征，因此采用降维方法排除贡献率较低的特征以提升预测性能。

参考文献

[1] Banerje, D. (2014) Forecasting of Indian Stock Market Using Time-Series ARIMA Model. 2014 2*nd International Conference on Business and Information Management* (ICBIM), Durgapur, India, 09-11 January 2014, 131-135. https://doi.org/10.1109/ICBIM.2014.6970973

[2] Rouf, N., Malik, M.B., Arif, T., Shar, S., Singh, S., Aich, S. and Kim, H.C. (2021) Stock Market Prediction Using Machine Learning Techniques: A Decade Survey on Methodologies, Recent Developments, and Future Directions. *Electronics*, **10**, Article No. 2717. https://doi.org/10.3390/electronics10212717

[3] Naz, N. and Reddy, Y.Y.R. (2023) Financial Applications of Machine Learning: A Literature Review. *Expert Systems with Applications*, **219**, Article ID: 119640. https://doi.org/10.1016/j.eswa.2023.119640

[4] Mustapa, F.H. and Ismail, M.T. (2019) Modelling and Forecasting S&P 500 Stock Prices Using Hybrid Arima-Garch Model. *Journal of Physics*: *Conference Series*, **1366**, Article ID: 012130. https://doi.org/10.1088/1742-6596/1366/1/012130

[5] Guo, W., Liu, Q., Luo, Z. and Tse, Y. (2022) Forecasts for International Financial Series with VMD Algorithms. *Journal of Asian Economics*, **80**, Article ID: 101458. https://doi.org/10.1016/j.asieco.2022.101458

[6] Chen, Y. and Hao, Y. (2017) A Feature Weighted Support Vector Machine and K-Nearest Neighbor Algorithm for Stock Market Indices Prediction. *Expert Systems with Applications*, **80**, 340-355. https://doi.org/10.1016/j.eswa.2017.02.044

[7] Khaidem, L., Saha, S. and Dey, S.R. (2016) Predicting the Direction of Stock Market Prices Using Random Forest. https://arxiv.org/abs/1605.00003

[8] Kim, Y. and Enke, D. (2016) Developing a Rule Change Trading System for the Futures Market Using Rough Set Analysis. *Expert Systems with Applications*, **59**, 165-173. https://doi.org/10.1016/j.eswa.2016.04.031

[9] Das, S.R., Mishra, D. and Rout, M. (2019) Stock Market Prediction Using Firefly Algorithm with Evolutionary Framework Optimized Feature Reduction for OSELM Method. *Expert Systems with Applications*: *X*, **4**, Article ID: 100016. https://doi.org/10.1016/j.eswax.2019.100016

[10] Srijiranon, K., Lertratanakham, Y. and Tanantong, T. (2022) A Hybrid Framework Using PCA, EMD and LSTM Methods for Stock Market Price Prediction with Sentiment Analysis. *Applied Sciences*, **12**, Article No. 10823. https://doi.org/10.3390/app122110823

[11] Row, S. (1997) EM Algorithms for PCA and SPCA. In: Jordan, M., Kearns, M. and Solla, S., eds., *Advances in Neural Information Processing Systems* 10, MIT Press, Cambridge, USA.

[12] Cordella, C.B.Y. (2012) PCA: The Basic Building Block of Chemometrics. *Analytical Chemistry*, **47**. https://doi.org/10.5772/51429

[13] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735