

基于ARIMA-CNN-LSTM的股票价格预测

董其成, 何利文

南京邮电大学物联网学院, 江苏 南京

收稿日期: 2024年9月18日; 录用日期: 2024年10月22日; 发布日期: 2024年10月31日

摘要

股票市场的波动会影响人们生活的各个方面, 因此准确预测股票价格具有重要意义。然而, 传统的时间序列预测模型(如ARIMA)在处理股票价格中的非线性特征时表现不佳, 难以获得令人满意的预测效果。鉴于深度学习在处理非线性问题上的优越能力, 它在股票价格预测中展现出巨大的潜力。本文提出了一种ARIMA-CNN-LSTM混合模型来预测股票价格, 该模型充分挖掘了股票市场的历史信息, 从而实现更高精度的预测。首先, 使用ARIMA对股票数据进行预处理, 将处理得到的残差序列输入CNN, 通过卷积提取股票数据的深层特征; 随后, 通过LSTM挖掘股票数据的长期时间序列特征。本文以中国银行2009年1月1日至2023年12月31日的股票数据为研究对象, 构建混合模型, 并与ARIMA-LSTM模型、CNN-LSTM模型的预测结果进行对比。实验结果表明, 本文提出的混合模型预测精度更高, 相较于ARIMA-LSTM与CNN-LSTM模型, 本模型在均方误差(MSE)指标上分别降低了25%与16.7%, 在平均绝对误差(MAE)指标上分别降低了14.1%与9.4%, 展示了更好的预测效果。

关键词

ARIMA, LSTM, 混合模型, 股票价格预测

Stock Price Prediction Based on ARIMA-CNN-LSTM

Qicheng Dong, Liwen He

School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing Jiangsu

Received: Sep. 18th, 2024; accepted: Oct. 22nd, 2024; published: Oct. 31st, 2024

Abstract

The volatility of the stock market affects various aspects of people's lives, making accurate stock price predictions highly significant. However, traditional time series forecasting models, such as ARIMA, perform poorly when dealing with the nonlinear characteristics of stock prices, leading to unsatisfactory prediction outcomes. Given the superior capability of deep learning in handling

nonlinear problems, it shows great potential in stock price prediction. This paper proposes a hybrid ARIMA-CNN-LSTM model for stock price prediction, which fully exploits historical market information to achieve higher prediction accuracy. First, ARIMA is used to preprocess the stock data, with the resulting residual series fed into a CNN, which extracts deep features of the stock data through convolution. Subsequently, LSTM is employed to capture the long-term temporal characteristics of the stock data. Using the stock data of the Bank of China from January 1, 2009, to December 31, 2023, as the research object, the hybrid model is constructed and compared with the prediction results of the ARIMA-LSTM model and CNN-LSTM model. The experimental results show that the proposed hybrid model achieves higher prediction accuracy. Compared to the ARIMA-LSTM and CNN-LSTM models, this model reduces the mean squared error (MSE) by 25% and 16.7%, respectively, and decreases the mean absolute error (MAE) by 14.1% and 9.4%, respectively, demonstrating better forecasting performance.

Keywords

ARIMA, LSTM, Hybrid Model, Stock Price Prediction

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

股市变化常被视为金融领域的“晴雨表”[1], 准确预测股票价格一直是投资者和金融机构关注的重点。然而, 由于股票市场的高度复杂性和不确定性, 精确预测股票价格始终是一个充满挑战的任务[2]。为了解决这些挑战, 研究人员提出了多种方法和模型, 以提升金融时间序列预测的准确性和效果。

传统时间序列预测模型, 如自回归移动平均模型(ARMA)和自回归移动平均模型(ARIMA), 已广泛应用于时序预测[3]。然而, 这些传统模型通常假设数据呈线性分布, 而金融数据往往具有非线性特征且存在长期记忆效应, 因此传统模型在金融领域的预测效果并不理想。随着深度学习的快速发展, 研究人员逐渐将其应用于金融市场[4]。卷积神经网络(CNN)、长短期神经网络(LSTM)、门控循环单元(GRU)等方法在金融预测方面成果显著。通过将这些深度学习模型与传统的时间序列模型结合, 能够在处理线性和非线性时序特征时提升预测精度。

因此, 本文提出了 ARIMA-CNN-LSTM 混合模型, 结合了 ARIMA 在处理线性趋势方面的优势与 CNN 和 LSTM 在处理非线性特征及长期依赖方面的能力。首先, ARIMA 用于去除时间序列中的线性成分和噪声, 随后通过 CNN 提取局部特征, LSTM 则负责捕捉长期依赖。通过这种分层次的处理, ARIMA-CNN-LSTM 能够全面建模金融时间序列中的复杂关系。为了详细阐述该模型的原理, 接下来将依次介绍 ARIMA、CNN 和 LSTM 的基础理论。

2. 基础理论

2.1. ARIMA

差分自回归移动平均(ARMA)模型和自回归综合移动平均(ARIMA)模型是经典的股票预测模型。自回归移动平均模型是自回归模型与移动平均模型的结合, ARMA (p, q)符合以下公式:

$$\alpha_t = \mu_0 + \sum_{i=1}^p \mu_i \alpha_{t-i} + \gamma_t + \sum_{i=1}^q \beta_i \gamma_{t-i} \quad (1)$$

公式 1 表示, 如果该序列是平稳的, 也就是说其行为不会随时间变化, 那么我们就可以根据该序列的过去行为来预测未来。ARIMA 也可以看作是自回归模型(AR)、滑动平均模型(MA)、差分(I)相结合[5], 使用 AR 捕捉时间序列的依赖性和趋势性, 使用 MA 捕捉时间序列数据的随机误差, 使用差分确保时间序列数据的稳定性。当序列是非平稳时, ARIMA(p, q, d)对序列采用和差分, 将序列变为平稳的。其中 p 是自回归项数, q 是移动平均项数, d 是将时间序列变为平稳状态所进行的差分次数。

2.2. CNN

卷积神经网络(CNN)是一种特殊的前馈神经网络, 其网络结构如图 1 所示, 组成包括卷积层、池化层、全连接层三部分[6]。CNN 的处理流程为: 首先通过卷积层提取输入数据特征, 然后使用池化层对提取的特征进行提炼, 最后在全连接层整合并输出数据。由于 CNN 采用了卷积运算, 相较于常规的矩阵运算, 其计算速度得到了显著提升。CNN 的卷积层和池化层的交替使用能够有效提取数据局部特征并降低局部特征维度[7]。

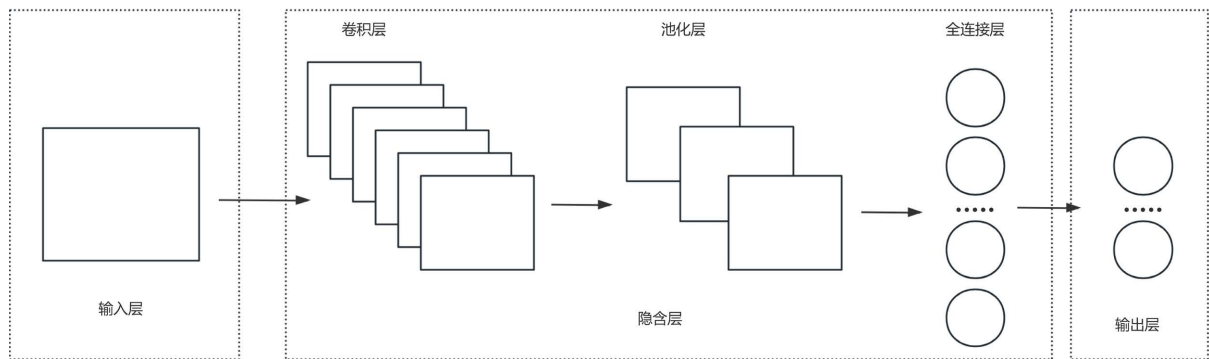


Figure 1. CNN network structure

图 1. CNN 网络结构

2.3. LSTM

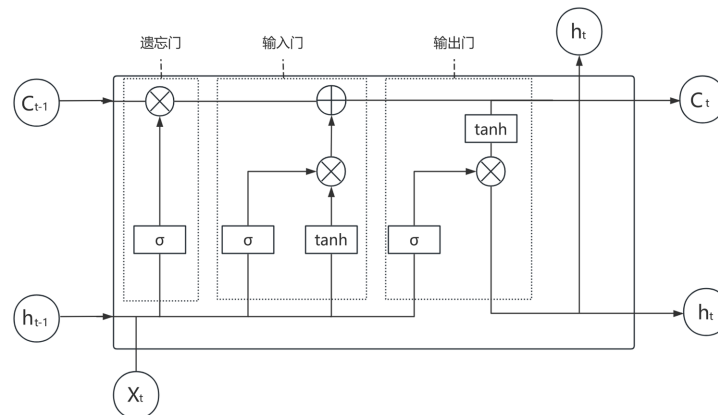


Figure 2. LSTM gating structure

图 2. LSTM 门控结构

长短期记忆神经网络(LSTM)是循环神经网络的一种, 它的出现是为了解决传统循环神经网络(RNN)在时序预测方面的缺陷, 即 RNN 难以处理具有长期依赖关系的数据, 会出现梯度消失问题[8]。LSTM 的关键在于借鉴了人脑选择性记忆和选择性遗忘的特点, 引入了遗忘门、输入门、输出门三个门控机制和

记忆单元[9]。通过门控机制, 不仅能够接受上层神经元的输出, 还可以选择性地保留有效历史信息。具体的门控结构如图 2 所示。

遗忘门决定从上一时刻记忆状态需要忘记什么信息, 本质上是通过 Sigmoid 函数, 对前一时刻的输出 h_{t-1} 和当前时刻的输入 X_t 进行操作, 生成 0~1 之间的值来代表遗忘多少信息。其中, 0 代表遗忘所有信息, 1 代表保留所有信息。

输入门可以决定哪些信息可以加入到上一时刻记忆状态并且更新为新的记忆状态, 本质上也是通过 Sigmoid 函数和 tanh 函数, 对 h_{t-1} 和 X_t 进行操作, 生成 0~1 之间的值来代表新信息的重要程度。

输出门决定了当前时刻可以输出的状态信息有哪些, 同样是通过 Sigmoid 函数对 h_{t-1} 和 X_t 进行操作, 再经过激活函数 tanh 映射, 生成 0~1 之间的值控制信息的输出。

3. ARIMA-CNN-LSTM 模型

ARIMA 擅长处理时间序列中的线性模式、趋势和季节性成分, 通过差分、移动平均和自回归, ARIMA 可以去除数据中的线性部分, 使序列更加平稳[10]。虽然 ARIMA 可以消除一部分随机噪声, 但它对非线性关系的处理能力有限, 不过可以通过它为后续深度学习模块提供了一个较为清晰的输入特征, 便于更高效地进行非线性建模。CNN 用于提取数据中的局部时序特征, 而 LSTM 则擅长捕捉时间序列中的长期依赖关系[11]。经过 ARIMA 预处理, 数据中的线性趋势被去除, 使得 CNN 和 LSTM 能够更好地处理复杂的非线性结构, 从而提升模型的整体性能。

本文使用的 ARIMA-CNN-LSTM 模型结构包括输入层、ARIMA 预处理层、CNN 层、LSTM 层和输出层。通过 ARIMA 对输入数据进行预处理, 去除数据中的线性趋势、周期性成分和随机噪声, 得到一个更为平稳的残差序列。残差序列是 ARIMA 未能解释的部分, 通常包含非线性特征, 适合输入深度学习模型进行进一步建模。经过 ARIMA 预处理, 原始序列中的大部分线性信息被去除, 保留了更加复杂的、深层次的非线性模式。

残差序列经过归一化后, 输入到 CNN 模块。CNN 通过多个卷积层提取时间序列中的局部特征。卷积层通过应用卷积核扫描序列数据, 捕捉短期依赖和局部特征, 通过这种操作识别时间序列中的短期特征并形成特征映射。形成的特征映射包含了 CNN 提取的局部特征, 能够有效反映数据中的短期依赖关系。这些映射将作为 LSTM 模块的输入。

LSTM 通过门控机制, 读取 CNN 输出的特征映射, 捕捉数据中的长期依赖关系, 进一步整合序列的时序特征。LSTM 不仅能够记住过去多个时间步的特征, 还可以忽略不重要的信息。LSTM 模块的输出是一个隐藏状态, 包含了对时间序列的全局时序特征的整合和处理结果。最后将 LSTM 层输出的隐藏状态传到输出层, 通过全连接层将 LSTM 的隐藏状态转化为预测结果并进行输出。

4. 数据处理与模型构建

4.1. 数据获取

本文对中国证券市场上的中国银行股价进行实证研究, 股票代码为 601988.SH, 数据来源为 Tushare, 股价数据在 Tushare 中会进行公开展示。选取了中国银行从 2009 年 01 月 01 日到 2023 年 12 月 31 日的数据进行研究, 共 3630 条数据(节假日除外), 包含日期、最高价、最低价、开盘价、收盘价、成交量、成交额等数据。

本实验将数据按 9:1 的比例划分为训练集和测试集, 训练集时间起止时间为 2019 年 01 月 05 日到 2022 年 07 月 05 日, 共 3266 条数据, 测试集的时间起止时间为 2022 年 07 月 06 日到 2023 年 12 月 29 日, 共 364 条数据。训练集和测试集的原始数据如图 3 所示。

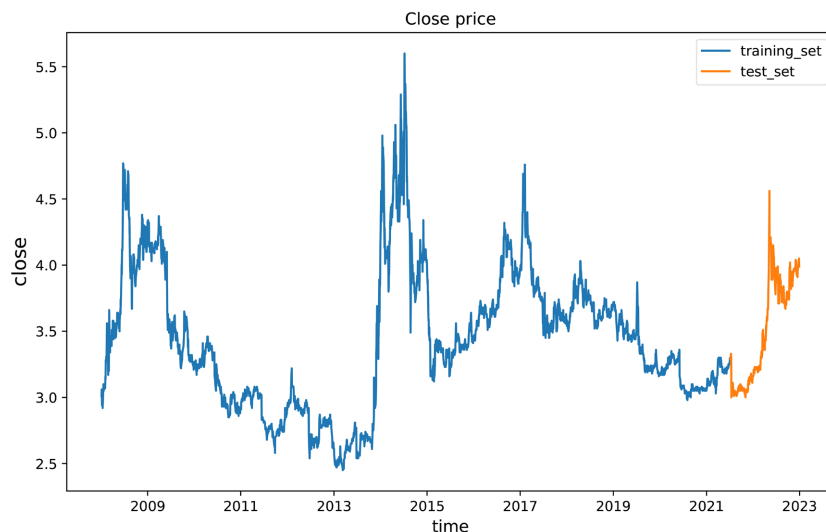


Figure 3. Training set and test set data
图 3. 训练集和测试集数据

4.2. ARIMA 模型构建

4.2.1. 数据平稳性校验

使用 ADF 检验来校验原始训练集的序列平稳性。验证结果如表 1 所示。

Table 1. The original sequence ADF verification result

表 1. 原始序列 ADF 校验结果

ADF 属性	ADF 属性值
Test Static Value	-2.84539
p-value	0.05208
Critical Value (1%)	-3.43237
Critical Value (5%)	-2.86343
Critical Value (10%)	-2.56725

如果一个时间序列是平稳的, 那么 Test Static Value 下方称为 t 值, 需要同时小于 1%临界值、5%临界值和 10%临界值[12]。由上表可以看出, t 值大于 10%临界值, 因此是不平稳的。使用 ARIMA 模型进行时序预测的数据要求是平稳的, 因此对原始训练数据进行一阶差分运算, 使数据变得平稳。一阶差分后的 ADF 检验结果如表 2 所示。从表中数据可以看出, t 值小于 1%临界值、5%临界值和 10%临界值, 因此是平稳序列, 可用于后续分析实验。

Table 2. The result of the first-order differential sequence ADF verification

表 2. 一阶差分序列 ADF 校验结果

ADF 属性	ADF 属性值
Test Static Value	-10.6038
p-value	6.0766e-19
Critical Value (1%)	-3.43237
Critical Value (5%)	-2.86343
Critical Value (10%)	-2.56725

4.2.2. 模型参数选择

通过 3.2 可以得到结论, 原始序列是非平稳的, 一阶差分序列是平稳的, $ARIMA(p, d, q)$ 模型中的差分次数 d 确定为 1, 下面需要确定 $ARIMA$ 模型中的 $AR(p)$ 和 $MA(q)$ 。使用自相关图(ACF)和偏自相关图(PACF)得到 p 值和 q 值。

ACF 和 PACF 如图 4 所示。其中 ACF 的横坐标表示滞后阶数, 即与当前时间点间隔的时间步数。纵坐标表示自相关系数, 反映了时间序列中不同滞后阶数下, 当前值与过去值之间的线性相关程度, 接近 1 表示正相关程度高, 接近 -1 表示负相关程度高, 接近 0 表示相关性弱。PACF 的横坐标同样表示滞后阶数, 纵坐标表示偏自相关系数, 反映的是在控制了中间滞后项的影响后, 特定滞后阶数下当前值与过去值之间的相关性。从图中可以看出, 在滞后阶数为 2 之后, PACF 处于基本稳定的状态, 因此确定 p 值为 2。对于任何的滞后阶数, ACF 都具有长尾, 因此采用 q 值为 0。此时, $ARIMA$ 模型的参数都已经确定, 即 $ARIMA(2, 1, 0)$ 。

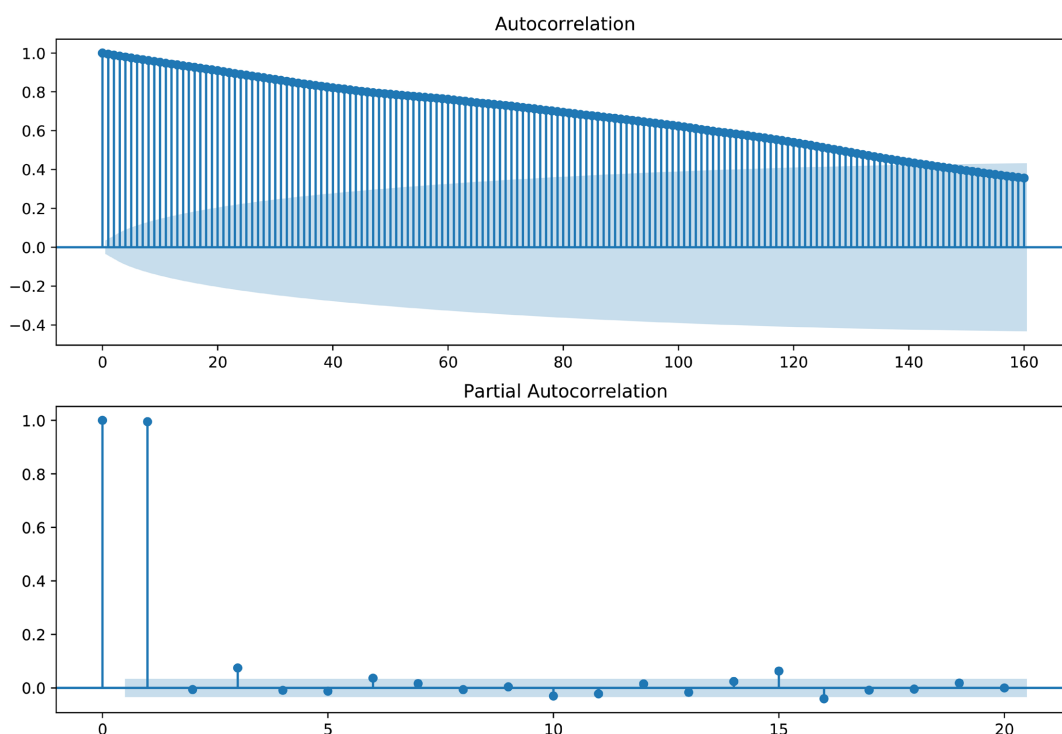


Figure 4. Autocorrelation plots and partial autocorrelation plots

图 4. 自相关图和偏自相关图

4.3. 混合模型构建

使用 CNN-LSTM 模型对 $ARIMA$ 模型生成的残差序列进行训练, 模型的建立过程如下。

步骤一: 数据预处理。使用 $ARIMA(2, 1, 0)$ 对原始序列进行处理, $ARIMA$ 模型会去除数据中的趋势和线性成分, 生成一个残差序列, 反映 $ARIMA$ 无法解释的非线性部分。为了提高后续模型训练的效果, 使用归一化将残差序列缩放到 $[0, 1]$ 范围内。将归一化后的残差序列划分为训练集和测试集, 其中 90% 作为训练集, 10% 作为测试集。

步骤二: 模型训练。将经过归一化处理的残差序列输入到 CNN 模块进行特征提取。将 CNN 的卷积核数量设为 32, 使用 ReLU 激活函数, 提升模型对非线性特征的捕捉能力。通过最大池化层降维并减少

特征冗余, 防止过拟合。将 CNN 提取的特征映射输入到 LSTM 模块, 用于捕捉时间序列中的长期依赖关系。设置 LSTM 层数为 5, 每层 LSTM 的隐藏单元大小为 64, 时间步长设置为 20, 在 LSTM 层中加入 Dropout, dropout 率设置为 0.3, 用以防止过拟合同时提高模型的泛化能力。为了可以加速模型收敛, 使用 Adam 对模型进行优化, 设置学习率为 0.01。其他参数, 将 epoch 设为 50, 这是初始的训练轮次。经过实验对比批次大小为 32 和 64 的效果, 选择 Batch Size 为 32, 在这个数值下, 模型有更好的收敛速度和稳定性。

步骤三: 模型训练结果。将 LSTM 输出结果进行反归一化, 将反归一化后的残差预测结果与 ARIMA 的线性拟合结果相加, 将 ARIMA 模型的线性部分和 CNN-LSTM 模型的非线性部分混合, 生成最终的预测结果。

4.4. 模型普适性测试

为了验证模型的普适性, 增加选择建设银行和工商银行的股票数据作为实验对象, 数据选取时间均为 2009 年 1 月 1 日至 2023 年 12 月 31 日, 其中 90% 的数据作为训练集, 10% 的数据作为测试集。与中国银行数据的处理方法一致, 采用相同的超参数来构建 ARIMA-CNN-LSTM 模型, 对数据进行预处理、模型训练和预测分析。首先, 使用 ARIMA 模型对每只股票的数据进行预处理, 去除数据中的线性趋势和季节性成分, 保留残差序列。接着, 残差序列经过 CNN 层提取局部特征, 通过卷积操作识别数据中的短期关系。在此基础上, LSTM 层捕捉时间序列中的长期依赖关系, 整合局部和全局的时序特征。

5. 实验结果与分析

5.1. 模型评价指标

为了评估 ARIMA-CNN-LSTM 模型在股票价格预测的准确性, 本文选取了均方误差(MSE)、均方根误差(RMSE)和平均绝对误差(MAE)三个指标来衡量混合模型在股票价格预测方面的性能。通过这三个指标可以反映实际值与预测值之间的误差大小, 三个指标的计算公式如下:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_{real} - y_{pred})^2 \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_{real} - y_{pred})^2} \quad (3)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_{real} - y_{pred}| \quad (4)$$

其中, y_{real} 表示 i 日股票的实际收盘价, y_{pred} 表示 i 日股票的预计收盘价, n 表示数据集的样本数量。

5.2. 不同模型预测结果对比

为了验证本模型在股票价格预测方面的效果, 本文选择了 ARIMA-LSTM 模型和 CNN-LSTM 模型作为基准模型, 将预测结果与 ARIMA-CNN-LSTM 模型进行对比。同时为了保证模型预测的公平性, 所有模型中的相同模块均设置相同的参数, 都使用 Adam 对模型进行优化。基于三种评价指标下的不同模型预测结果如表 3 所示, 对比中国银行在不同模型的预测误差结果, 可以看出 ARIMA-CNN-LSTM 模型三种指标数值都是最低, 在股票价格预测方面的准确性有所提升。

通过对建设银行和工商银行股票数据的预测结果进行分析, 可以发现 ARIMA-CNN-LSTM 模型在这两个数据集上同样表现出色。MSE、RMSE 和 MAE 指标显示, 模型在三个银行数据上的预测误差显著低

于 ARIMA-LSTM 和 CNN-LSTM 模型, 表明 ARIMA 预处理和 CNN 局部特征提取提升了模型的整体表现。

Table 3. Comparison of prediction results of different data in different models
表 3. 不同数据在不同模型预测结果比较

股票名称	模型	MSE	RMSE	MAE
中国银行	ARIMA-LSTM	0.0020	0.0447	0.0248
	CNN-LSTM	0.0018	0.0424	0.0235
	ARIMA-CNN-LSTM	0.0015	0.0387	0.0213
建设银行	ARIMA-LSTM	0.0084	0.0916	0.0552
	CNN-LSTM	0.0079	0.0889	0.0524
	ARIMA-CNN-LSTM	0.0065	0.0806	0.0482
工商银行	ARIMA-LSTM	0.0049	0.0700	0.0465
	CNN-LSTM	0.0046	0.0678	0.0439
	ARIMA-CNN-LSTM	0.0038	0.0616	0.0386

6. 结论

为了提高股票价格的预测精度, 提出了 ARIMA-CNN-LSTM 的混合模型, 以 2009 年 1 月到 2022 年 7 月中国银行的股价为参考, 预测 2022 年 7 月到 2023 年底的收盘价, 经过实验对比分析, 可以得到以下结论。

(1) 相比于其他深度学习的混合模型, 本文提出的 ARIMA-CNN-LSTM 模型具有更小的预测误差, 能够更准确地预测股票等时间序列的未来走势。通过比较 ARIMA-LSTM 模型和 ARIMA-CNN-LSTM 模型的预测结果, 可以看出 CNN 在提取局部特征时, 对模型的预测精度有显著的提升作用。在 ARIMA-LSTM 模型中加入 CNN 后, ARIMA-CNN-LSTM 模型能够提取时间序列中的局部模式, 使得残差序列的非线性特征被更好地捕捉, 从而提高了整体的预测能力。

(2) ARIMA-CNN-LSTM 模型中, ARIMA 的预处理步骤有效地去除了数据中的线性趋势和季节性成分, 这为后续的 CNN-LSTM 模块处理非线性特征提供了更干净的输入数据。通过对比 CNN-LSTM 模型和 ARIMA-CNN-LSTM 模型的预测结果可以看出, 先使用 ARIMA 进行线性预处理, 再通过 CNN-LSTM 模块处理残余的非线性关系, 能够有效提升模型的整体预测精度。

(3) 通过对中国银行、建设银行和工商银行股票数据的实验验证, 证明了 ARIMA-CNN-LSTM 模型在不同股票数据上的普适性和鲁棒性。该模型不仅能够有效处理线性趋势和非线性波动, 还能够通过局部和长期特征提取提高预测精度。这表明 ARIMA-CNN-LSTM 模型在金融市场预测中具有广泛的应用潜力。

综上, 本文提出的 ARIMA-CNN-LSTM 混合模型, 相比于实验中使用的其他模型, 在减少误差方面有一定的优势, 可以预测更准确的股票价格。

参考文献

[1] 张苏林, 宋迦南. 基于 CNN-LSTM 模型股票价格预测研究[J]. 合作经济与科技, 2023(3): 58-62.
[2] 朱子敬, 何利文. 基于 ABC-LSTM-GRU 的时间序列分解与预测模型[J]. 软件工程, 2024, 27(3): 58-62.
[3] 肖田田. 基于 K-means-LSTM 模型的证券股价预测[J]. 科技和产业, 2024, 24(3): 210-215.
[4] 齐太威, 于文年. 基于 LSTM 的多指标股票预测[J]. 计算机与数字工程, 2024, 52(2): 337-342.

-
- [5] 龙宇, 许浩然, 余华云, 等. 基于 ARIMA-LSTM-XGBoost 组合模型的铁路货运量预测[J]. 科学技术与工程, 2023, 23(25): 10879-10886.
- [6] 李靖萱. 基于 CNN-BiLSTM 和注意力机制的汇率预测研究[D]: [硕士学位论文]. 福州: 福建师范大学, 2023.
- [7] 王世明, 张少童, 娄嘉奕. 基于 CNN-LSTM-ARIMA 的超短期风速预测研究[J/OL]. 新能源进展, 1-9.
https://kns.cnki.net/kcms2/article/abstract?v=UjEBX92ALNEWSGB7mnaCviYl-qC1UI_PCJj4Z-bpP8ocwXi-JaLswVmTjaXH759pZgarUkTlc0cF87Zbj3xtZMtYE20xg1HtMoY5UUwuYmTx-oBtzaB4E4Cg4EQmZRH7yHOJ5oBBMERRwcov-GCGZvfXrlnjmw7xwX37zdUXef0vBMvqnv4OqZjJqj1oKG9GV&uniplatform=NZKPT&language=CHS, 2024-01-09.
- [8] Singh, P., Jha, M., Sharaf, M., El-Meligy, M.A. and Gadekallu, T.R. (2023) Harnessing a Hybrid CNN-LSTM Model for Portfolio Performance: A Case Study on Stock Selection and Optimization. *IEEE Access*, **11**, 104000-104015.
<https://doi.org/10.1109/access.2023.3317953>
- [9] Zhang, J., Ye, L. and Lai, Y. (2023) Stock Price Prediction Using CNN-BiLSTM-Attention Model. *Mathematics*, **11**, Article No. 1985. <https://doi.org/10.3390/math11091985>
- [10] 王瑞, 李瑞沂, 曹沛根, 等. 基于 ARIMA-LSTM 混合模型对传染病的预测分析[J]. 现代信息科技, 2024, 8(1): 116-120.
- [11] 梅杰. 基于二维 GoogLeNet-LSTM 模型对股票涨跌的预测[D]: [硕士学位论文]. 上海: 上海师范大学, 2023.
- [12] Shi, Z., Hu, Y., Mo, G., et al. (2022) Attention-Based CNN-LSTM and XGBoost Hybrid Model for Stock Prediction. arXiv: 2204.02623.