Advances in Social Sciences 社会科学前沿, 2022, 11(11), 4517-4527
Published Online November 2022 in Hans. http://www.hanspub.org/journal/ass
https://doi.org/10.12677/ass.2022.1111617

Hans汉斯

基于LSTM模型的股价预测研究

——以保利发展为例

古龙飞，王　鑫

成都锦城学院，电子信息学院，四川 成都

收稿日期：2022年8月24日；录用日期：2022年10月25日；发布日期：2022年11月2日

摘　要

股票市场受到国际形势、市场行情和国家政策等多因子影响，简单的K线图与机器学习预测很难抗击股价的波动随机性，导致股价预测精度不能达到投资参考水平。本文采用LSTM神经网络模型仿真保利发展(600048) 2018年2月13日至2022年3月30的时间序列收盘价，通过调节隐含单元数进行多次仿真，得到误差较小、可行性较强的结果。分析实验结果得到，向后预测的收盘价数据，得到的误差比极低。因此，本文实验得到LSTM模型能对长期随机波动的时间序列数据做出预测，且模型的精度高、可推广性强。

关键词

股票市场，K线图，LSTM，时间序列数据

Stock Price Forecasting Research Based on LSTM Model

—Taking Poly Development as an Example

Longfei Gu, Xin Wang

Department of Electronic Information Engineering, Chengdu Jincheng College, Chengdu Sichuan

Received: Aug. 24th, 2022; accepted: Oct. 25th, 2022; published: Nov. 2nd, 2022

Abstract

The stock market is affected by multiple factors such as international situation, market conditions and national policies. It is difficult for simple K-graph and machine learning prediction to resist the randomness of stock price fluctuations so that the accuracy of stock price prediction cannot

文章引用: 古龙飞, 王鑫. 基于 LSTM 模型的股价预测研究[J]. 社会科学前沿, 2022, 11(11): 4517-4527.
DOI: 10.12677/ass.2022.1111617

reach the investment reference level. This paper uses LSTM neural network model to simulate the time series closing price of Poly Development (600048) from 13 February 2018 to 30 March 2022. By adjusting the hidden unit number for many times, the results of small error and strong feasibility are obtained, and by analyzing the closing price data, the error ratio is very low. Therefore, the LSTM model can predict the time-series data with long-term random fluctuations, with high accuracy and strong generalizability.

Keywords

Stock Market, K Line Chart, LSTM, Time Series Data

1. 引言

股票市场预测的研究对于投资者以及企业具有重要意义，对于推动我国经济发展发挥重要作用，股票价格预测模型发展趋势一直备受股民关注，然而因股价数据存在波动性、非线性、受市场影响随机变化等多方面因素，导致对股价模拟与向后预测存在较大的困难。随着国民投资知识与意识的不断增强以及大数据、人工智能在金融投资界的广泛应用，利用人工智能算法、机器学习以及深度学习等方法进行股票市场预测成为当前热门研究话题，一个好的模型对股价进行模拟能够为投资者提供明确的投资指导，为投资者和金融集团有效规避风险，带来更高的回报率。其中神经网络(Neural Networks, NN)模型是一种模拟人大脑的非线性动力学习系统，在股票市场预测方面，能够进行有效高精度拟合性的预测，神经网络根据对以往股票变化特征数据的学习，拟合出股票价格变化的规律，最终从一定程度上反映未来市场的股票波动趋势。国内外相关学者提出通过长短记忆神经网络(Long Short-Term Memory, LSTM)模型来对金融股票市场波动进行定量预测的效果最佳，本文旨在通过 LSTM 模型对保利发展时序收盘价数据进行模拟，分析不同隐含层个数组合下模型预测股价精确度，验证模型的可行性与有效性。

2. 文献综述

股票市场中随机变化较多，股民在股票市场中投资以收益率高、风险低作为首要目的，如何利用技术手段进行股价分析规避风险已引起投资者和专家的重视。在目前计算机技术与人工智能技术迅速应用的背景下，有许多学者开始建立数学模型进行股价的预测对选股策略进行探究。

股价是一组无序波动的时序数据，陈博闻(2021)建立 ARIMA 模型结合 R 语言对时间序列数据进行预测分析，结果得到 ARIMA 模型处理短期时间序列有较好的仿真效果，不适用于大量的历史数据模拟[1]。针对股价具有波动性的特点，王东(2021)引入 PCA 对数据进行降维处理，再将降维后的数据与股票相关 KDJ、MACD 指标数据一同输入 LSTM 神经网络模型，避免了数据信息重叠、输入信息变量多、泛化性差的问题，优化了输入股价数据，得到了更为精确的股票预测效果[2]。在模型创新上，文宝石(2020)针对股价具有非线性等特点，提出经数据多维处理后的 LSTM 组合模型，实验结果表明多维处理后的组合模型误差更小[3]。席小雅(2022)将百度股票作为研究对象，利用 4195 个收盘价长期单列时序数据输入 LSTM 神经网络模型进行股价预测，对模型参数进行调整优化后，再结合 5 个股票指标进行数据处理，建立多特征 LSTM 模型，得到多特征模型拟合效果更好，且能更为准确地预测股价数据变化趋势的结论

[4]。梁宇佳(2021)另辟蹊径地从主观方面思考，将机器学习应用于情感分析中，结合情感分析结果建立LSTM 机器学习组合模型，得到了优于传统模型的实验结果[5]。除此之外，自股价预测成为热门话题后，许多学者从多方面入手进行分析，冯宇旭(2019)基于沪深 300 指数对 LSTM、SVR、Adaboost 模型预测股价效果进行对比评估，分析第二日预测结果得到 LSTM 较优于其他两个模型。接着将多种模型合成然后进行岭回归，一方面吸取各模型的优势更加精准地预测了股价，另一方面为股民买卖决策提供了新的参考信息[6]。陈伟斌(2020)比较分析 ARMA、GM(1,1)、BPNN、SVR、LSTM 5 种模型在股价预测上的效果情况，并且对比分析得到 LSTM 拥有长短期记忆，相较于其他模型能更精准稳健地预测股价[7]。黄超斌(2021)基于上证综合指数数据将 BP、CNN、RNN、GRU、LSTM 神经网络模型股票价格预测结果进行对比分析，发现 LSTM 神经网络的 MSE 值最低，预测值和真实值的拟合程度最高，且 LSTM 模型继承了 RNN 处理时序数据的能力，并自身具有长期记忆的优点可以避免对历史数据进行过度依赖，有更好的股价预测效果[8]。

综上所述，已有的研究中显示 LSTM 模型预测股价能够利用模型长期记忆的特点得到较为良好的股价仿真效果。本文沿用 LSTM 神经网络模型预测保利发展的收盘价数据，通过调试不同隐含层单元数探究模型的预测精度，利用 Matlab 建立数学模型得到不同隐含层个数不同的仿真效果，对比实际曲线和真实曲线走势判定模型的精度，分析模型的 RMSE 和向后预测收盘价误差比得到模型的可行性较强、精度较高，可用于股价预测的推广。

3. 指标选取

1) 数据采集

在指标确定分析过程中，考虑到股票走势时由多因子联合决定，为直观反映股价走势且精确地预测未来几天股价走势，本文以股价的收盘价数据作为研究对象。除此之外，文章为追求研究具有代表性，选用房地产行业较大企业保利发展(600048)作为模型仿真对象，利用同花顺官方网站下载了保利发展 2018 年 2 月 13 日~2022 年 3 月 30 日收盘价数据，总计 1000 个精确数据，收盘价数据格式如下表 1 所示。

Table 1. Partial data of Poly Development closing price
表 1. 保利发展收盘价部分数据

日期	收盘价	日期	收盘价
2018-02-13	15.11	2022-03-17	15.44
2018-02-14	15.2	2022-03-18	16.98
2018-02-22	15.43	2022-03-21	16.67
2018-02-23	15.64	2022-03-22	17.22
2018-02-26	15.09	2022-03-23	17.23
2018-02-27	14.51	2022-03-24	17.07
2018-02-28	14.63	2022-03-25	16.77
2018-03-01	14.7	2022-03-28	16.94
2018-03-02	14.8	2022-03-29	16.8
2018-03-05	14.86	2022-03-30	17.72

2) 数据预处理

本文模型构建利用上千个时间序列收盘价数据，在下载大量收盘价数据时，获得的数据可能会存在

噪声，该噪声会直接影响收盘价仿真和预测精度，本文在建立模型前先对数据进行去噪处理，再利用Matlab 对数据缺失值、异常值进行填补。除此之外，收盘价是一个随机波动、难以精准预测走势的数据，很容易出现不稳定的情况，首先对收盘价数据进行差分处理，可以提高数据稳定性，再将处理后的数据进行模型仿真可以提高模型的精确度。

3) 股票预测基本原理

已知影响股价变动的因素众多，通过将变动因素和历史数据设置成合理参数，分割为训练集、验证集、测试集，进行训练得到模型，构建模型之后再通过调整参数，变换结构，加入算法等方式来优化模型，基于该模型进行股价的仿真，分析预测曲线额实际走势曲线评估模型的精度。一种方法是通过比对预测数据和真实数据两者的绘图，若算法模拟得出的股价预测图像与实际股价走势图像吻合或者近似，表明模型预测股价有较好的精度，反之则说明该模型还存在进步空间。另一种方式是通过预测数据的误差累计值来计算平均误差率，从而得到该模型的预测精度水平，模型的均方根误差值越小则仿真结果越好，该值越大则说明该模型还需要改进。

4. 模型原理极其构建

1) 模型原理

随着机器语言、互联网技术的发展，神经网络在不同层面上都得到了补足性、创新性的变革。在发展上，神经网络衍生出了 RNN、CNN、BP 等各式各样的神经网络模型；在应用上，神经网络现被广泛的运用于计算机、金融、医药等行业。其中，RNN 神经网络能够解决时序、语音、文本等问题，但却面临着一些有待解决的问题：RNN 神经网络在反向传播过程中存在指数级较大的梯度，会导致网络崩溃，有造成模型梯度爆炸的潜在危机；RNN 神经网络在网络层数较多时，经过正向传播和反向传播的过程后会使权重的影响变得很小，导致模型有梯度消失的可能性；RNN 神经网络只有一个细胞状态，当时间序列数据较长时，网络很难对之前的信号保持有效的记忆状态，所以导致了网络只有短期依赖、只能记忆短期输入时序数据等问题。而由 Hochreiter 和 Schmidhuber 等人提出的 LSTM 模型能有效解决的问题[9]，LSTM 神经网络的核心就是在 RNN 神经网络循环结构的基础上引入 σ 函数，从而由一个细胞状态变为两个细胞状态，隐含层有了更为复杂的相互作用，有效地解决了 RNN 模型梯度爆炸、梯度消失、短期记忆的问题，对股票价格的预测有较高的研究意义。

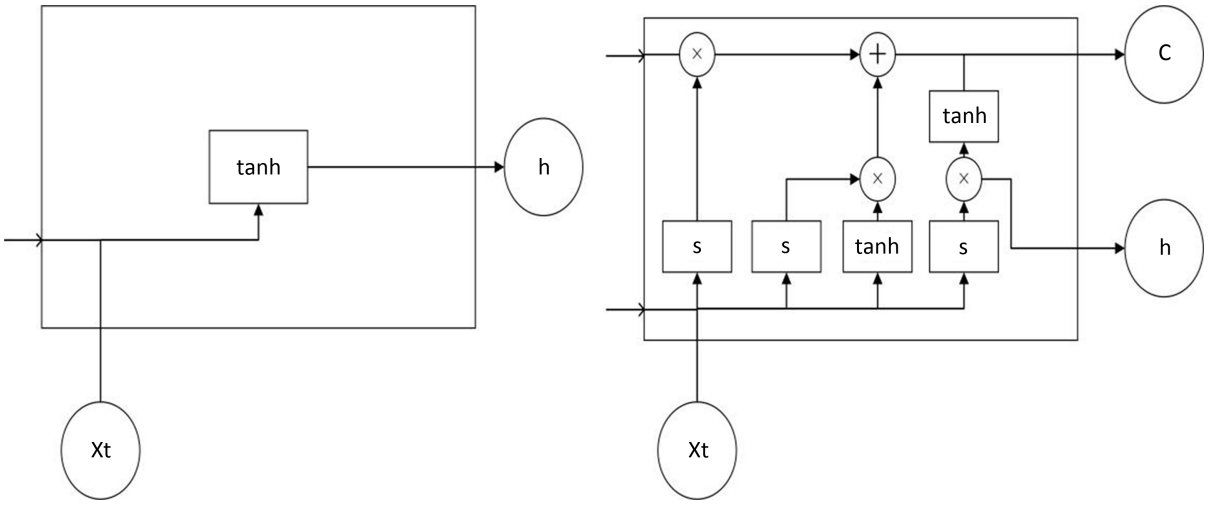

Figure 1. Schematic diagram of the difference between RNN and LSTM
图 1. RNN 和 LSTM 区别示意图

由上图 1 可以看出 LSTM 的基本结构有记忆单元、遗忘门、输入门、输出门等，每一部分都发挥着重要的作用，是模型中必不可少的基石。为进一步的将 LSTM 运用在股票价格预测上，需要清楚模型的基本原理和工作机制，现以 t 时刻为研究时点，探讨经过 LSTM 神经网络之后，细胞状态的变化，信息遗忘、更新、记忆的具体演算过程：

首先，是遗忘门将信息选择性遗忘的过程，其工作如图 2 所示。LSTM 神经网络需要通过 σ 函数确定哪些由 x_t 输入的信息需要遗忘，哪些不需要遗忘。其中 σ 层输出值的取值范围是[0,1]，越靠近 0 表明信息遗忘的越多、越靠近 1 表明信息遗忘的越少，等于 1 时所有的信息都能保留，等于 0 时所有的信息都被舍弃。

$$f_t = \sigma\left(w_f \times \left[h_{t-1}, x_t\right] + b_f\right) \tag{1}$$

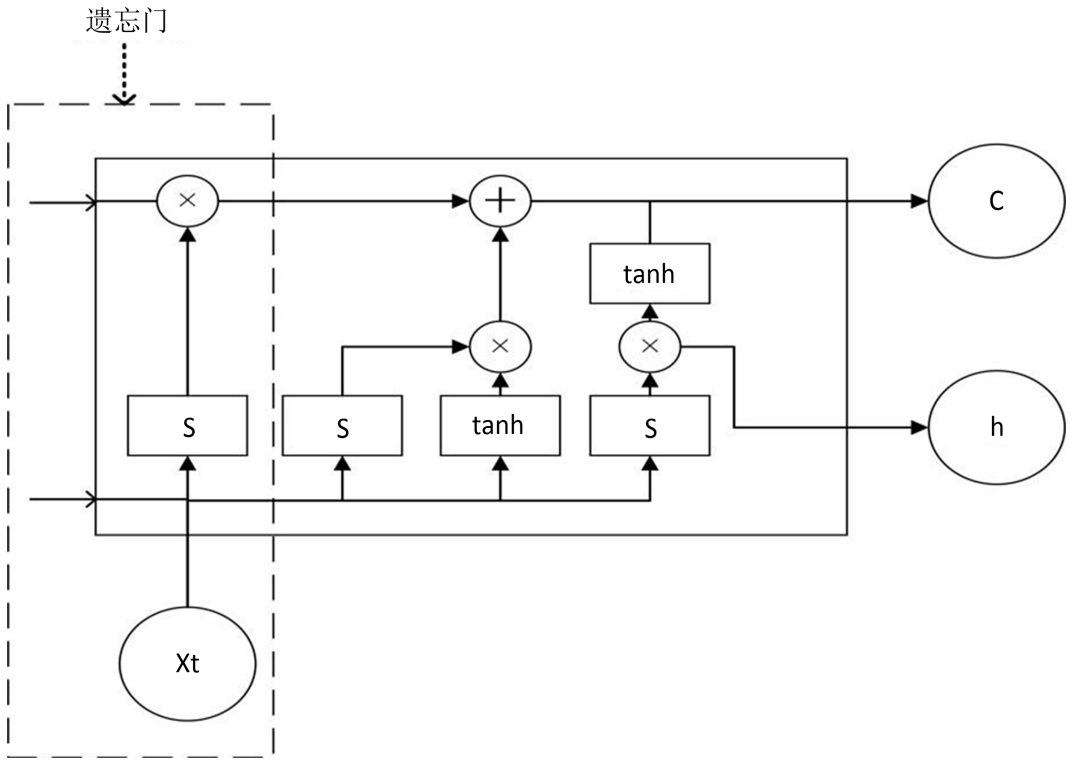

Figure 2. Forgetting gate process
图 2. 遗忘门过程

其次，是输入门录入并校正参数的更新过程，其工作如图 3 所示。整个更新过程分为两个步骤，第一步是 sigmoid 层在隐藏层输入信息的基础上通过 σ 函数确定更新哪些信息，第二步是 tanh 层通过 tanh 激活函数创建一个新的候选值向量。tanh 函数调节流经神经网络的信息将其值控制在[−1,1]之间。接着通过综合两步骤，确定细胞状态的变化情况以及信息的更新状况。其中输出的 C_t 作为细胞状态管理着长期记忆。

$$i_t = \sigma\left(w_i \times \left[h_{t-1}, x_t\right] + b_i\right) \tag{2}$$

$$\tilde{C}_t = \tanh\left(w_c \times \left[h_{t-1}, x_t\right] + b_c\right) \tag{3}$$

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \tag{4}$$

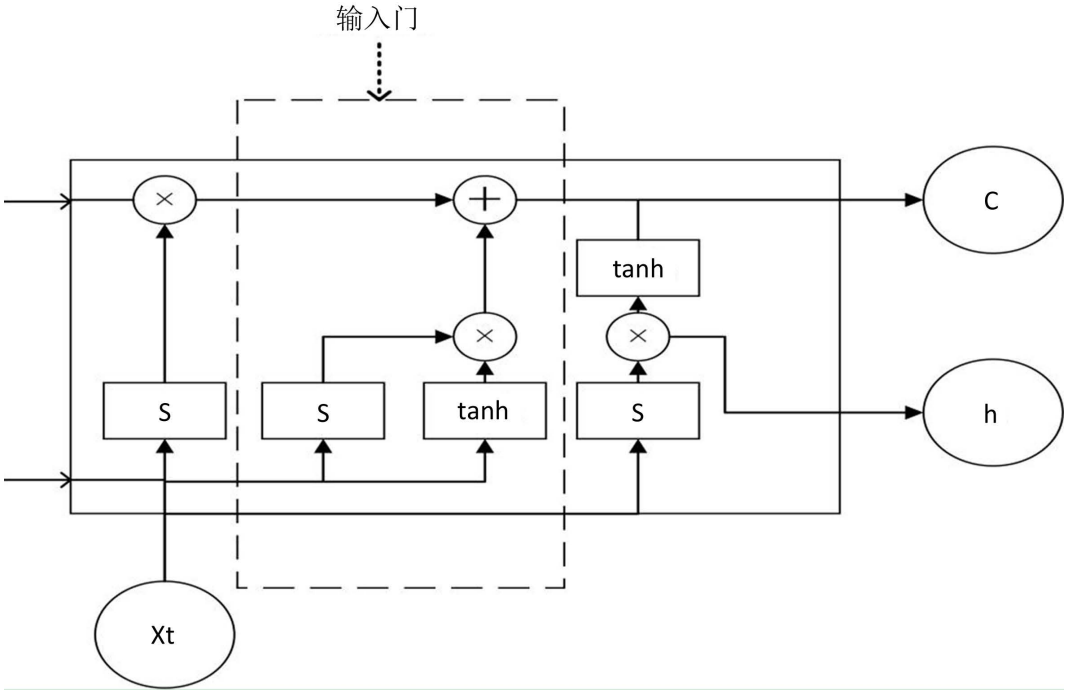

Figure 3. Input gate process
图 3. 输入门过程

再者，是输出门输出和校正参数的输出过程，如图 4 所示。输出门能够决定哪些信息可以输出，凭借其选择输出性，能够精准地控制信息的流出比例。隐藏层的输入状态通过 sigmoid 函数转换后与由 tanh 层函数控制的[−1,1]之间的值相乘后，输出最终的信息 h_t，其中输出的 h_t 作为输出状态存放着短期记忆。

$$o_t = \sigma\left(w_o \times \left[h_{t-1}, x_t\right] + b_o\right) \tag{5}$$

$$h_t = o_t \times \tanh\left(C_t\right) \tag{6}$$

以上公式中：f_t、i_t、o_t 分别是 t 时刻遗忘、输入、输出门的控制参数；w_f、w_i、w_c、w_o 分别表示 t 时刻遗忘、输入、输出门和单元状态的权重；$\tilde{C}_t$ 是 t 时刻细胞状态的候选向量；b_f、b_i、b_c、b_o 分别表示遗忘、输入、输出门和单元状态的偏置向量；σ 和 tanh 分别表示 sigmoid 与 tanh 激活函数；x_t 和 h_t 分别为 t 时刻的输入、输出状态。

2) LSTM 模型构建

结合 LSTM 神经网络长依赖性、长期记忆等特点，以 LSTM 神经网络为核心构建股票价格预测时间序列框架，本文将整个预测过程分为 3 个部分，分别是输入层、以 LSTM 神经网络为核心的隐藏层、输出层，框架结构示意图如下图 5。

输入层：收集保利地产 600048 股票从 2018/2/13~2022/3/30 的收盘价时序数据信息作为原始股票的时间序列，即模型的测试集，接着经过数据处理软件进行归一化和标准化的处理，消除不同量纲，将最终的结果作为隐藏层的输入值。

隐藏层：股票每一时点的输入层信息都会对 LSTM 神经网络造成影响，模型会进行输入信息的选择性遗忘，信息的更新，信息的输出过程，最终每个时点的状态值会传输到输出层。

输出层：输出层会将隐藏层传入的信息进行股票预测原理的运算过程，再通过反标准化处理，以便获取与原始股票预测时间序列在同一维度的预测结果。

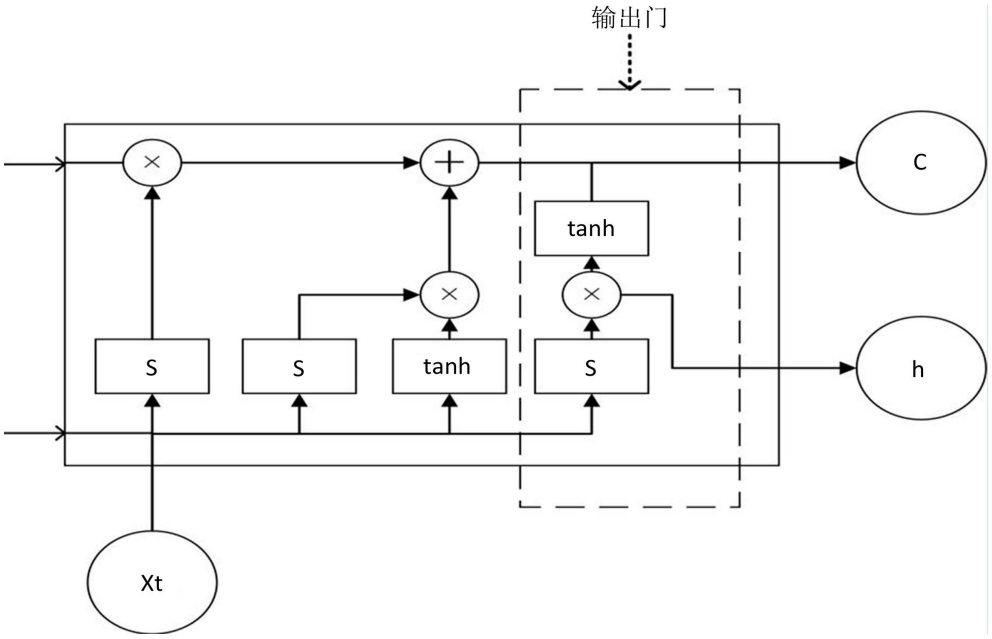

Figure 4. Output gate process
图 4. 输出门过程

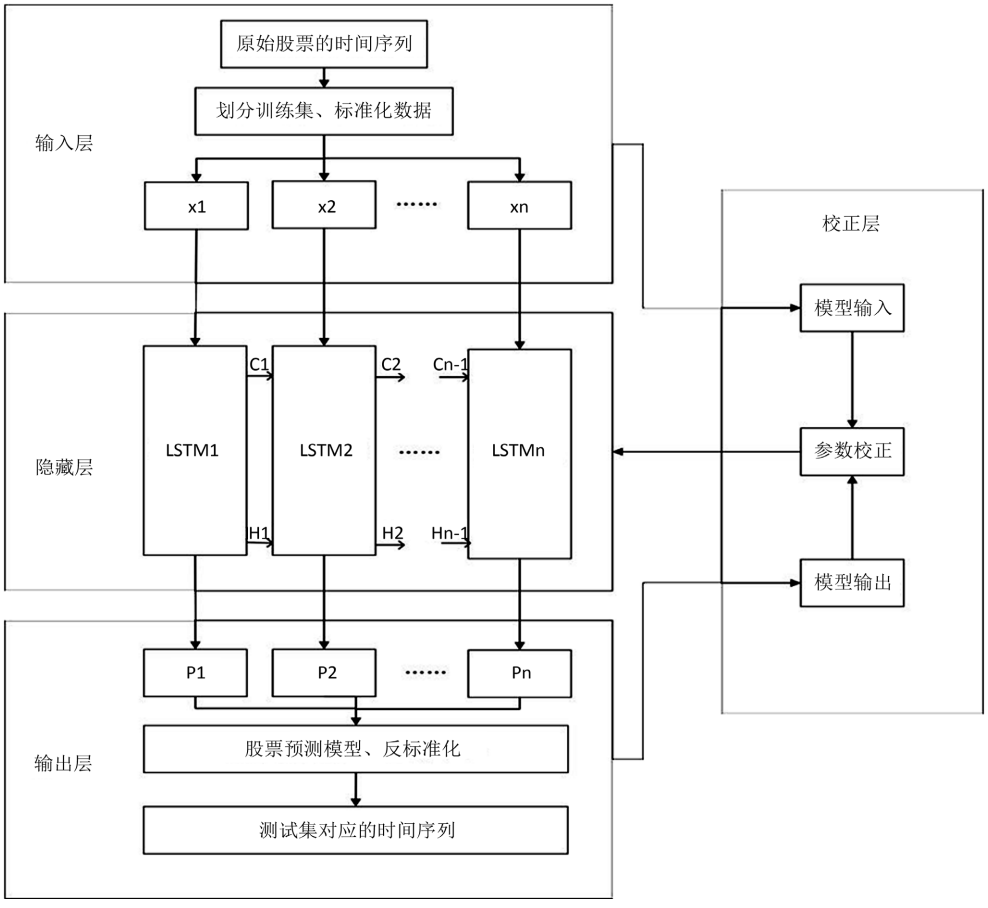

Figure 5. Block diagram of stock price prediction based on LSTM neural network
图 5. 基于 LSTM 神经网络的股票价格预测框图

校正层：在实际预测的过程中还要考虑神经网络的参数设置是否合理并应据此进行校正，因此还需要设置一系列的测试集对神经网络进行反馈，以优化神经网络，达到股票预测的最优参数设置状况。

5. 仿真极其结果分析

1) 算法仿真

本文将采集的 1000 个收盘价数据先进行数据预处理后导入 Matlab 中构建 LSTM 模型，把 1000 个时序收盘价数据作为模型的输入指标，把前 90%即 2018-2-13 到 2021-11-2 的数据作为训练集，将后 10%即 2021-11-3 到 2022-3-30 的收盘价数据作为检验集，将收集的收盘价数据仿真如图 6 所示。建立 LSTM 模型过程中，收盘价输入为 1 维，则输出为 1 维，初始学习率此模型设置为 0.005，迭代设置为 200 次。本文通过调整隐含层数量和节点数调整预测精度，隐含层数量为 2 或 3，节点数为 64、128 或 256，调整隐含单元个数组合数量进行多轮仿真模拟得到不同的预测结果，其实验效果见表 2。

从图 5 和表 2 可以得到，LSTM 神经网络模型仿真收盘价 1000 个输入数据有较好的效果，隐含单元数不同模型有不同的仿真效果，算法运行时间和 RMSE 都随参数不同而改变，要获得仿真时长最短、均方根误差最小的隐含单元参数需要进行多次仿真进行调试。

2) 结果分析

有相关研究表明，LSTM 进行长期时序数据预测时，隐含层数为 2 或 3 时仿真结果最为精确，同时节点数设置为 2 的 N 次方，通过调整 N 的大小和隐含层数量得到不同的预测精度，本文进行 LSTM 仿真得出了不同的实验结果，仿真结果如图 7 和图 8 所示。

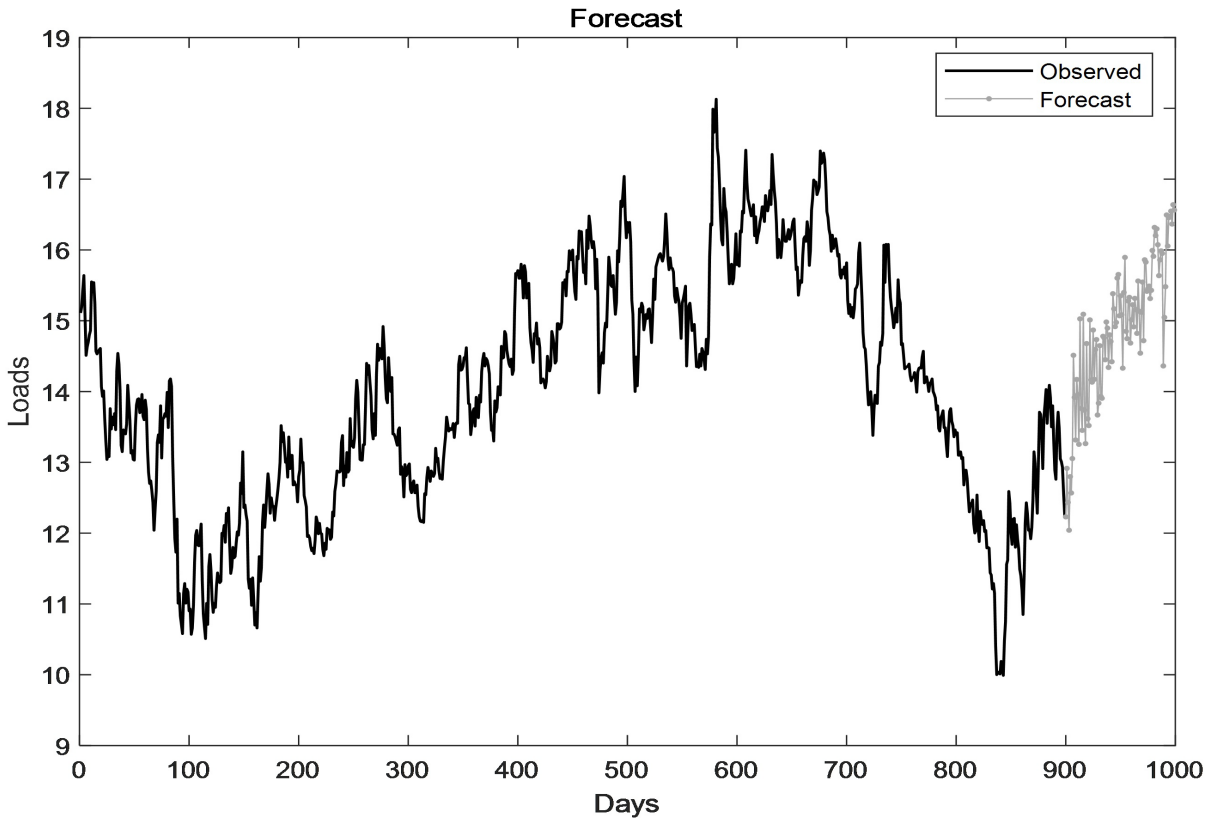

Figure 6. Simulation diagram of 1000 input data
图 6. 1000 个输入数据仿真图

Table 2. Simulation results of parameter adjustment model
表 2. 参数调整模型仿真结果

隐含单元个数	仿真时长	RMSE
2*64	12 s	0.4422
2*128	15 s	0.34684
2*256	29 s	0.4185
3*64	12 s	0.6083
3*128	33 s	0.36403
3*256	78 s	0.41566

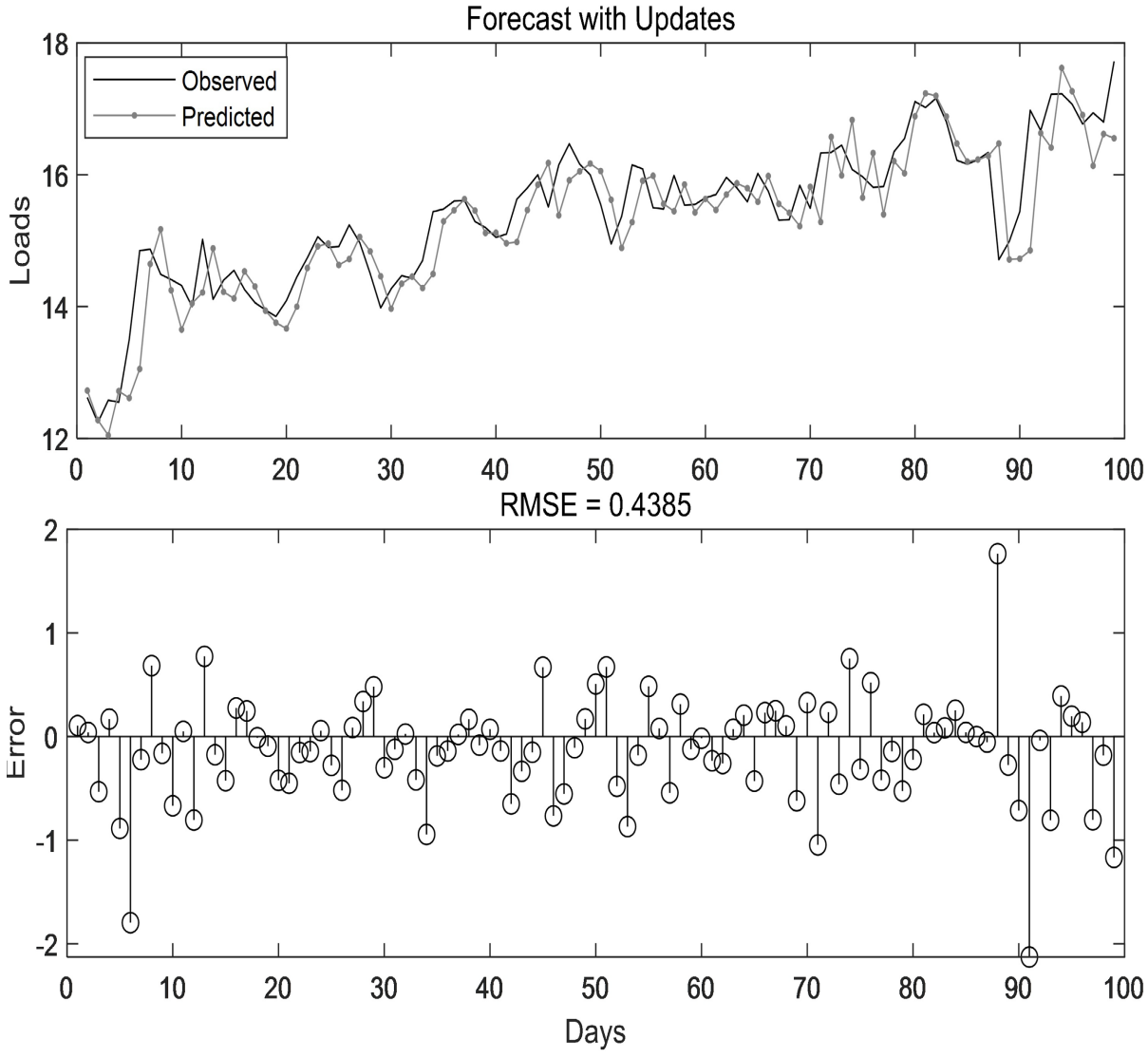

Figure 7. Simulation results at 2*256
图 7. 2*256 时仿真结果

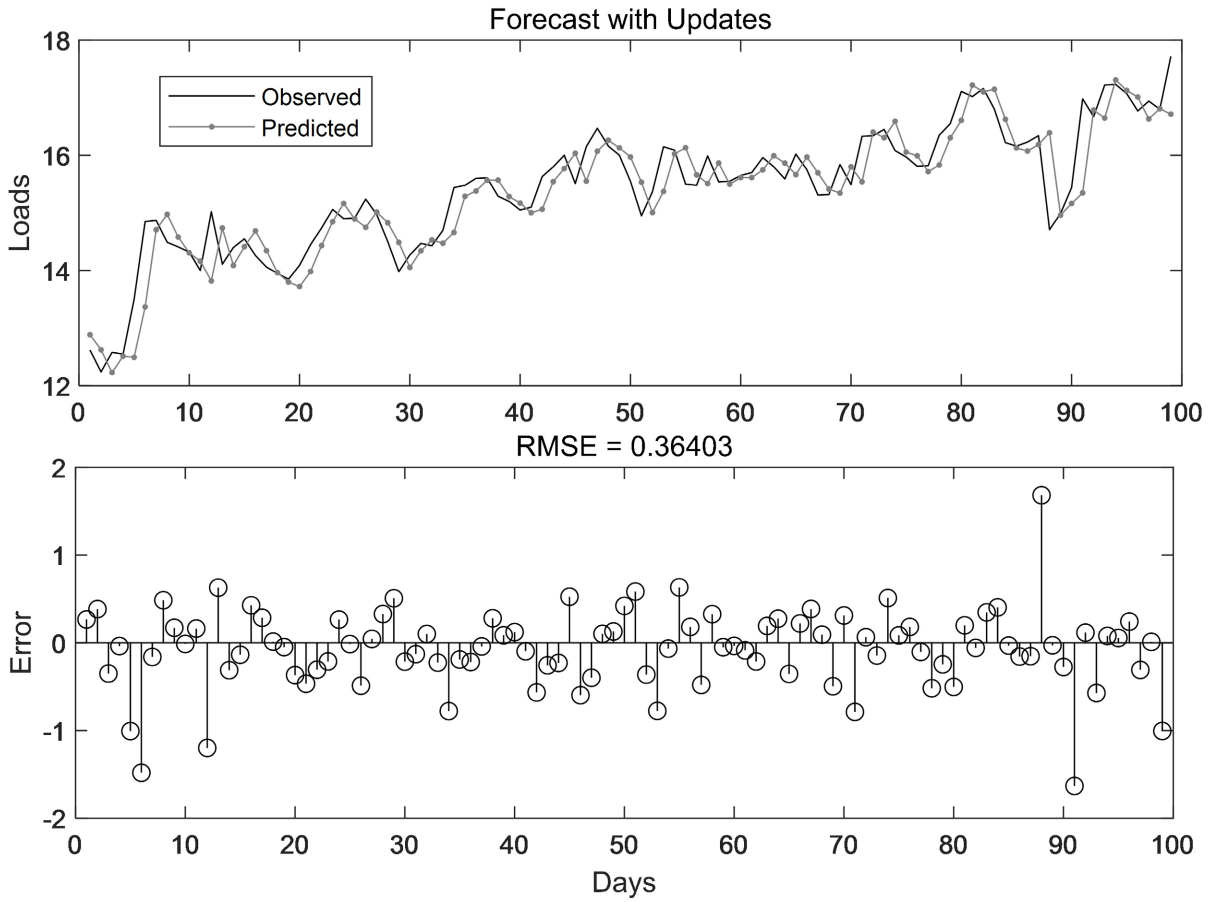

Figure 8. 3*128 simulation results
图 8. 3*128 仿真结果

输入数据后 10%用于检验得到曲线如上所示，图中信息可以清晰得到，LSTM 神经网络模型对于收盘价数据模拟效果较为精确，收盘价时序数据仿真后得到的预测曲线(灰色)走势和真实值(黑色)走势基本吻合，由此可以得到 LSTM 神经网络模型处理时序收盘价数据时性能较好。模型的预测性能指标 RMSE(均方根误差)数值能够直接表明算法的性能和可行性，均方根误差公式如下：

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\left(X_{\mathrm{Prediction},t}-X_{\mathrm{Observed},t}\right)^{2}} \tag{7}$$

其中，N 代表数据数量，$X_{\mathrm{Prediction},t}$ 表示预测值，$X_{\mathrm{Observed},t}$ 表示实际值。模型仿真时微调隐含单元个数进行最好预测效果模拟，2*256 与 3*128 所配置的参数不同，得到的拟合曲线走势偏差不大，但是得到的 RMSE 可以得到，2*256 相较于 3*128 出现了过拟合的现象，LSTM 神经网络算法对股价进行预测时需要严格控制隐含单元数，数量过小，得到的 RMSE 可能会较大，是因为此时算法处于欠拟合的状态，但并不是隐含单元数越大越好，当隐含单元数过大时，仿真得到的 RMSE 也可能出现较大的现象，是因为模型已经处于是一个过拟合的状态，此时误差会较大。因此，进行 LSTM 算法时，必须进行多次调试控制隐含单元数，才能得到 RMSE 较小、拟合速度较快的参数值。

3) 误差分析

调整参数完成后，本文进行了算法的仿真，还利用保利发展 2018 年 2 月 13 日~2022 年 3 月 30 日收盘价数据进行了向后预测，得到了未来五个工作日的股价预测数据，数据结果如表 3 所示。

Table 3. LSTM backward forecast 5-day stock price data
表 3. LSTM 向后预测 5 天股价数据

日期	预测值	实际值	误差比
2022/3/31	17.48	17.7	0.0124
2022/4/1	18.39	18.21	0.0098
2022/4/6	18.54	18.3	0.0131
2022/4/7	17.72	17.62	0.0056
2022/4/8	18.43	18.21	0.121

分析图表可以得到，利用前 1000 个历史数据预测未来的 5 个股价数据，预测值与实际值虽然不是完全一样，但是得到的预测值和实际值误差相差不大，经过计算后得到的误差比基本在 1%左右，表明 LSTM 神经网络模型利用时序数据预测未来数据有较好的精确度，在股价预测方面能够进行推广。

6. 结语

股票市场预测的研究受到了投资者、商业界和学术界的广泛注视，基于深度学习理论，本文分析股票预测基本原理并构建 LSTM 模型对保利发展公司的历史数据进行模拟。利用 2022 年 3 月 30 日前 1000 个工作日时序收盘价数据作为输入建立 LSTM 模型，通过调节隐含单元个数对时序数据进行多次实验仿真，分析仿真结果得到，在不同的隐含单元数组合中，LSTM 模型对收盘价数据模拟效果较为精确，预测曲线和实际曲线走势基本相同，误差较小，由此得到本文应用的 LSTM 模型预测股价可以呈现出较好的预测效果。同时，分析向后预测 5 个工作日的收盘价数据直观得到，预测值和实际值存在一定的误差，但最终的误差比较小，因此实验结果表明该模型可以用于股价预测中，且有较高的精度和可行性。由于总体模型的预测精度仍存在进步空间，在后续的研究中可加入其他预测模型以完善该模型不足之处，或加入语言文字信息、市场指数中的相关特征对模型进行训练以提高预测的精准度与科学性，希望为投资者和学者提供更加科学合理的理论化参考。

参考文献

[1] 陈博闻. 基于技术指标及 ARIMA 模型预测股票价格——以中国平安保险集团公司股票调整后的收盘价为例[J]. 统计与管理, 2021, 36(7): 53-57.

[2] 王东, 王霄鹏, 杨川东. 一种基于主成分 LSTM 模型在股票预测中的研究[J]. 重庆理工大学学报(自然科学), 2021, 35(2): 282-288.

[3] 文宝石, 颜七笙. 数据多维处理 LSTM 股票价格预测模型[J]. 江西科学, 2020, 38(4): 443-449+472.

[4] 席小雅, 秦荷斌, 鲁志娟. 基于 LSTM 神经网络模型的股票价格变化预测研究——以百度股价为例[J]. 全国流通经济, 2022(16): 102-105.

[5] 梁宇佳, 宋东峰. 基于 LSTM 和情感分析的股票预测[J]. 科技与创新, 2021(21): 126-127.

[6] 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): 308-315.

[7] 陈伟斌, 林奕真, 王宗跃. 股票信息挖掘与 LSTM 预测[J]. 集美大学学报(自然科学版), 2020, 25(5): 385-391.

[8] 黄超斌, 程希明. 基于 LSTM 神经网络的股票价格预测研究[J]. 北京信息科技大学学报(自然科学版), 2021, 36(1): 79-83.

[9] 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