

基于 LSTM 神经网络的股票开盘价预测研究

陈欢

成都理工大学

DOI:10.32629/ej.v9i5.3529

[摘要] 股价序列具有非线性特征与长期记忆性,使得以ARIMA为代表的传统时序模型在预测精度上存在明显局限。本文选取万科、平安银行、建设银行、凯盛科技四只A股,尝试以LSTM神经网络对开盘价进行预测。数据区间为2017—2025年,涵盖开盘价、成交量等9项指标,经Min-Max归一化和滑动窗口构造样本后,输入4层LSTM网络训练。结果显示,训练轮次设为600时模型收敛较优,凯盛科技的RMSE为0.3131、MAPE为0.51%,建设银行RMSE为0.2491。模型在平稳行情下拟合效果良好,但对极端冲击的响应存在滞后。上述结果可为量化投资与风险管控提供一定参考。

[关键词] LSTM; 股票开盘价预测; 时间序列; 深度学习

中图分类号: F272.1 **文献标识码:** A

Prediction of Stock Opening Price Based on LSTM Neural Network

Huan Chen

Chengdu University of Technology

[Abstract] Stock price series exhibit pronounced nonlinear dynamics and long-memory properties, rendering conventional time-series models such as ARIMA inadequate for capturing persistent dependencies. This study examines four A-share stocks—Vanke, Ping An Bank, China Construction Bank, and Triumph Science and Technology—and applies Long Short-Term Memory (LSTM) networks to forecast opening prices. Using historical trading data from 2017 to 2025 with nine indicators including opening price and trading volume, we construct samples through Min-Max normalization and an 8-day sliding window, then feed them into a four-layer LSTM. Empirical results suggest that 600 training epochs yield satisfactory convergence: Triumph Science and Technology achieves an RMSE of 0.3131 and a MAPE of 0.51%, while China Construction Bank records an RMSE of 0.2491. The model tracks price movements reasonably well under stable market conditions, yet exhibits noticeable lag in response to extreme shocks. This study provides a reference technical framework for quantitative investment and risk management.

[Key words] LSTM; Stock Opening Price Prediction; Time Series; Deep Learning

1 引言

金融市场作为现代经济的核心,其波动性与复杂性对全球经济格局产生着深远影响。股票价格易受宏观政策调控、投资者预期等多因素的影响,表现出高度非线性与动态演化特征^[1]。以ARIMA和GARCH为代表的传统时序方法依赖线性假设和固定参数结构,难以捕捉股票价格序列中的长期依赖关系^[2]。

近年来,深度学习技术的发展为金融预测提供了新思路。其中LSTM凭借其特有的门控机制,能够有效处理时间序列中的长距离依赖问题,在金融预测领域展现出显著潜力^[3]。该网络内部的记忆单元可通过遗忘门与输入门的动态调节,能够选择性保留或更新历史信息,从而实现对股价跨期关联模式的捕捉。现有研究多聚焦于股票收盘价或综合指数的预测,而针对于开盘价

作为体现隔夜信息吸收程度及开盘时段市场情绪的重要变量,相应的预测模型仍需进一步探究^{[7][8]}。

本文以万科、平安银行、建设银行、凯盛科技四只A股为研究对象,基于LSTM构建股票开盘价预测模型,主要工作包括:数据预处理与特征工程、LSTM模型构建与参数优化、预测结果分析与评估。

2 LSTM神经网络原理

2.1 LSTM基本概念

传统循环神经网络(RNN)在处理长序列时容易遭遇梯度消失问题,并且当时间步长增加时,较早时刻的信息对后续输出的贡献会呈指数级衰减^[4]。而长短期记忆神经网络(LSTM)作为RNN的一种改良结构,引入了遗忘门、输入门和输出门三种调控机制,

实现对信息传递的动态管理,有效解决了梯度消失问题^[4]。

2.2 LSTM核心计算流程

LSTM的核心计算包括以下几个步骤:

(1)遗忘门:负责决定上一时刻的细胞状态中 C_{t-1} 中有多少信息需要被丢弃。该门通过Sigmoid函数对输入信息进行筛选,输出值越接近0表示信息越应被遗忘,越接近1则表示信息应被保留。其计算形式为:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

(2)输入门与候选值:用于控制当前时间步的新信息以何种比例加入记忆单元。候选状态 $\tilde{C}_t$ 经由tanh激活函数生成,代表当前输入可能带来的新信息。两者的计算公式分别为:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

(3)细胞状态更新:在遗忘门与输入门的协同作用下,上一时刻的细胞状态 C_{t-1} 被有选择地遗忘,同时新的候选信息 $\tilde{C}_t$ 被整合进来。更新规则为:

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

(4)输出门:决定当前细胞状态 C_t 中有多少信息被传递到隐藏状态 h_t 。首先通过Sigmoid函数计算输出门的值 o_t ,再与经过tanh处理的细胞状态相乘,得到最终的隐藏状态:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t \odot \tanh(C_t)$$

上述各式中, x_t 为当前时间步的输入向量, h_t 为隐藏状态, C_t 为细胞状态, σ 为Sigmoid激活函数,tanh为双曲正切激活函数, $\odot$ 表示逐元素相乘。

3 数据预处理与模型构建

3.1 数据来源

实验数据选取平安银行(000001)、万科集团(000002)、建设银行(601988)、凯盛科技(600552)四只A股,数据区间为2017到2025年,通过筛选后四只股的有效样本量分别为1992、1990、1994、1900条。研究的指标因子包括股票的开盘价、最高价、最低价、收盘价、成交量、成交额等共9项数据特征。

3.2 数据预处理

针对数据中的缺失值,实验中采用线性插补方式进行填补,以保障时序数据的完整性。为了消除不同指标之间的量纲影响,对特征变量进行Min-Max标准化处理,将数据映射到[0, 1]区间,计算公式为:

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

采用滑动窗口方法构建时序特征矩阵,设定时间步长为8天,即利用以前8天的特征序列来预测第9天的开盘价。按8:2比例划分训练集与测试集。

3.3 模型架构设计

本研究的LSTM预测网络共堆叠4个隐藏层,各层神经元数分别为64、64、128、128,其层间插入Dropout层防止过拟合,且丢

弃率设为0.4^[5]。输入层接收时序特征矩阵(时间步长=8,特征数=8),输出端通过全连接层映射得到开盘价的预测值。模型训练采用均方误差(MSE)作为损失函数,Adam优化器更新参数^{[5][9]}。

4 实验结果与分析

4.1 评价指标

本文采用均方根误差(RMSE)和平均绝对百分比误差(MAPE)作为模型性能评价指标。

均方根误差计算公式为:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{t=1}^N (y_t - \hat{y}_t)^2}$$

平均绝对百分比误差计算公式为:

$$\text{MAPE} = \frac{1}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\%$$

4.2 不同训练周期的对比

实验对比了训练周期(Epoch)为50、100、200、400、600时的模型性能。结果表明,随着训练周期增加,模型损失值快速下降并从第10个周期起趋于平稳。当训练周期为600时模型收敛效果最佳,测试集预测误差最小。

4.3 四只股票的预测结果

四只股票的开盘价预测误差如下表所示:

企业	均方根误差(RMSE)	平均绝对百分比误差(MAPE)
平安银行	0.7911	1.03%
万科集团	3.4683	2.25%
建设银行	0.2491	1.33%
凯盛科技	0.3131	0.51%

从上表中可以看出,建设银行和凯盛科技的预测误差比较小,前者RMSE为0.2491、MAPE为1.33%,后者RMSE为0.3131、MAPE为0.51%,则预测得比较准确。而万科集团的均方根误差和平均绝对百分比误差明显比其他企业的大,万科是地产类股票,推测可能与2020年以来受到的疫情冲击、2021年的“三条红线”政策调控等外部环境有关,影响了其股票价格序列的平稳性,造成波动性大。

4.4 结果分析

结合本实验数据可得,该模型在一定程度上是有效的,数据预测与真实数据的拟合效果还是比较好的,能够跟踪短期价格走势并识别短期波动特征^{[6][10]},但长期来看,面对价格大幅波动时偏差明显,说明模型在捕捉长期趋势和极端波动方面存在不足,对极端市场的适应能力较差,难以应对价格的大幅快速变动。

5 结论

本文基于LSTM构建了预测模型,选取万科、平安银行、建设银行、凯盛科技四只A股来验证模型的有效性,主要结论如下: LSTM对金融时序数据中的非线性模式具有一定提取能力,建设银行与凯盛科技的预测结果(RMSE分别为0.2491、0.3131)误差小比较准确,可以证明其有效性;通过Min-Max归一化、滑

动窗口长度及训练轮次的合理设置, 可以提高预测精度, 提升性能; 模型在金融、科技等低波动行业预测误差显著低于地产等政策敏感行业, 以本文的万科为例, 其预测偏差显著高于其他的股票。

后续可考虑引入宏观经济指标、行业特征等外部变量作为影响因子来训练模型, 同时把历史极端事件数据构建训练集, 以增强模型的鲁棒性和泛化能力, 再者结合注意力机制等技术提升模型的可解释性, 推动模型向高鲁棒性、强可解释性升级, 为智能投顾、风险管控等场景提供更可靠的技术支撑。

[参考文献]

- [1]张尧庭.金融数据的统计分析[M].北京:科学出版社,2008.
- [2]王宏禹.时间序列分析——参数统计方法[M].北京:科学出版社,2010.
- [3]孙志军,薛磊,许阳明,等.深度学习研究综述[J].计算机应用研究,2012,29(8):2806-2810.

[4]HochreiterS.,SchmidhuberJ.,Longshort-termmemory[J].NeuralComputation,1997,9:1735-1780.

[5]王雪华,董文永.基于LSTM循环神经网络的短期电力负荷预测[J].电力系统保护与控制,2017,45(14):65-71.

[6]李航.统计学习方法[M].北京:清华大学出版社,2019.

[7]周志华.机器学习[M].北京:清华大学出版社,2016.

[8]GoodfellowI.,Bengio Y.,Courville A.Deep Learning[M].Cambridge:MIT Press,2016.

[9]吴微,陈维强,刘波.基于LSTM的股票价格预测研究[J].计算机工程与应用,2018,54(20):126-131.

[10]张贵勇,张春霞.改进的LSTM在金融时间序列预测中的应用[J].计算机科学,2019,46(11):231-236.

作者简介:

陈欢(2005--),女,汉族,四川南充人,本科,研究方向为金融数据分析与深度学习。