Advances in Applied Mathematics 应用数学进展, 2022, 11(7), 4887-4899
Published Online July 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.117513

Hans汉斯

基于多步特征选择的股市预测方法

张　娜

兰州交通大学，数理学院，甘肃　兰州

收稿日期：2022年6月22日；录用日期：2022年7月18日；发布日期：2022年7月25日

摘　要

智能金融预测模型在投资决策中发挥着重要作用。为解决高维数据引起的维数灾难和过拟合问题，有效提高多元非线性金融时序预测精度，设计了一种基于多步特征选择的GA-LSTM预测框架。首先，通过基于过滤和嵌入的特征选择方法从大量指标中筛选有效股票影响因子，然后将剩余变量输入遗传算法优化后的长短期记忆神经网络来预测股票收盘价。为了验证所提模型的有效性，将此模型与传统降维模型PCA、LASSO和预测模型ARIMA、MLP、SVR、RNN、GRU对比分析，在中国银行数据集上的实验结果表明：基于机器学习新的组合方法不仅可以大规模降维，预测误差MSE、MAPE也低于对比模型，显著提高了预测精度。最后，将此模型应用在中国不同行业代表性的股票中取得较好预测效果，再次证明此模型在金融时间序列智能特征提取和预测上的应用价值。

关键词

特征选择，深度学习，遗传算法优化，个股预测，长短期记忆网络

Chinese Stock Prediction Method Based on Multi-Step Feature Selection

Na Zhang

School of Mathematics and Physics, Lanzhou Jiaotong University, Lanzhou Gansu

Received: Jun. 22nd, 2022; accepted: Jul. 18th, 2022; published: Jul. 25th, 2022

Abstract

Intelligent financial forecasting model plays an important role in investment decision-making. In order to solve the problem of dimensional disaster and overfitting caused by high-dimensional data and effectively improve the prediction accuracy of multivariate nonlinear financial time se-

文章引用: 张娜. 基于多步特征选择的股市预测方法[J]. 应用数学进展, 2022, 11(7): 4887-4899.
DOI: 10.12677/aam.2022.117513

ries, a GA-LSTM prediction framework based on multi-step feature selection is designed. First, the effective stock impact factors are screened from a large number of indicators by the feature selection method based on filtering and embedding, and then the remaining variables are input into the long-short-term memory neural network optimized by the genetic algorithm to predict the stock closing price. In order to verify the effectiveness of the proposed model, this model is compared with traditional dimensionality reduction models PCA, LASSO and prediction models ARIMA, MLP, SVR, RNN, and GRU. The combined method can not only reduce the dimensionality on a large scale, but also the prediction error MSE and MAPE are lower than those of the comparison model, which significantly improves the prediction accuracy. Finally, this model is applied to the representative stocks of different industries in China and achieves good forecasting effect, which proves the application value of this model in intelligent feature extraction and forecasting of financial time series again.

Keywords

Feature Selection, Deep Learning, Genetic Algorithm Optimization, Stock Prediction, Long Short-Term Memory Network

1. 引言

随着金融市场的不断完善，金融时间序列数据量日趋庞大且波动更为复杂。股票作为一种重要的金融工具，它的变化不仅受投资者关注，也关乎企业甚至国家的经济发展。股票价格受到时间序列的随机性[1]、宏观政策、企业经济运行状况等诸多因素的影响，影响因素间复杂的非线性动态交互关系使序列存在非线性、非平稳、低信噪比的特点[2]。近年来，智能金融预测可以极大提高金融风险管理和相关决策分析的效率。在金融预测过程中，一些特征对模型的贡献几乎为零，不仅会限制结果准确性，还会占用机器学习模型更多的训练时间从而导致过拟合。因此，在考虑尽可能多的因素下，对高维数据降维，并进一步预测股票走向，为投资者决策提供数据支持仍然是学术界和金融界具有极大挑战的问题。因此，本文利用多用多步特征选择对影响因子降维，再结合深度学习方法预测股票收盘价。

2. 股票预测模型

本文建立的股票预测模型框架流程图如图 1 所示。首先，利用 Python 的 Tushare 接口爬取股票数据，并使用 TA-Lib 库计算技术指标。其次，对收集到的数据进行预处理。然后，用多步特征选择筛选变量产生特征子集作为预测模型的输入变量。再将筛选出的最优特征变量导入遗传算法优化的 LSTM 神经网络，对个股的收盘价进行预测，并计算预测误差。最后对比所提模型与多种方法的预测误差。

2.1. 多步特征选择

市场条件、公司自身、行业政策等因素会影响股票价格。对于股票分析一般是股票基本面和技术分析两部分。技术分析是通过编译一系列技术指标来完成的。本文选取我国各行业企业股票走势作为研究对象。以收盘价为输出变量，输入变量包括 K 线基本指标、成交量指标、价格指标、动量指标和波动率指标[3]，如表 1 所示。

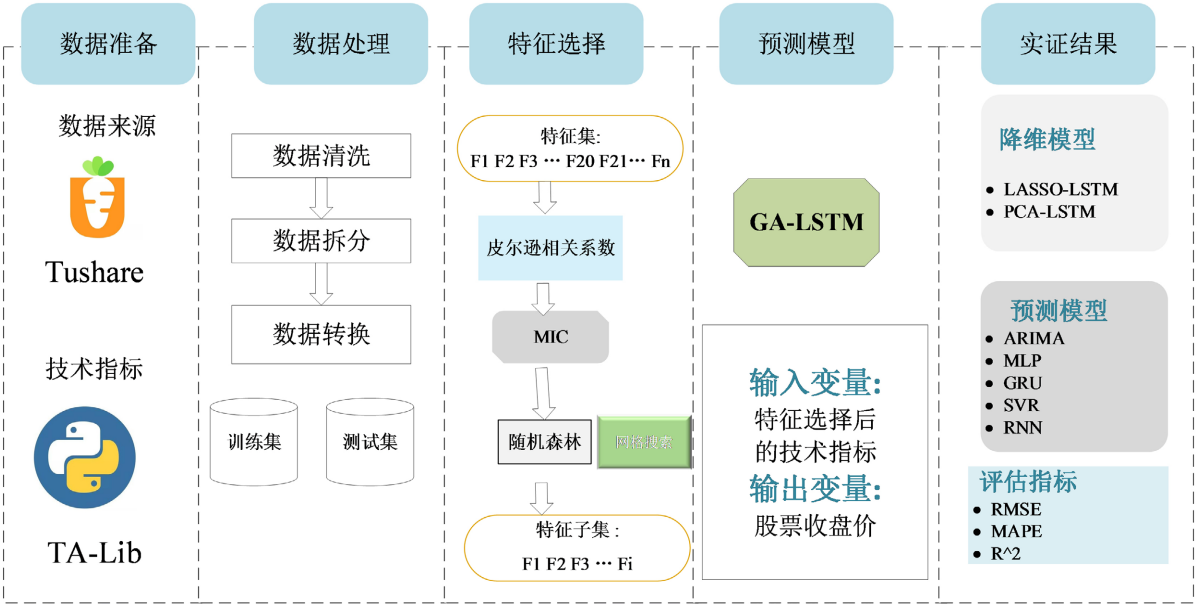

Figure 1. Proposed model procedure for stock prediction
图 1. 所提模型股票预测流程图

Table 1. Technical indicators of stocks
表 1. 股票技术指标

指标类型	指标名称
量价指标	能量指标(CR)、成交率比率及其 M 日简单移动平均(VR、MAVR、vr_6_smadx_6_ema)、振荡指标(WR(6)、WR(10))、随机指标(KDJ)、平行线差指标(DMA)、能量潮指标(OBV)、上升下降方向线(PDI、MDI)、乖离率(BIAS)
动量指标	平均趋向指数及其 M 日均值(ADX、ADXR)、APO、顺势指标(CCI)、动量震荡指标(CMO)、DX、指数平滑异同移动平均线(MACD)、资金流量指数(MFI)
波动率指标	动量指标(MOM)、价格震荡指标(PPO)、变化率(ROC)、相对强弱指标(RSI(6,12day)、rs_6,9,12)、中长线(TRIX)、终极波动(UOS)、威廉指标(WILLR)、middle_14_sma、动向指标(dm)、真实波动幅度及其均值(TR、ATR)、上升方向线(pdi_14、pdi)、下降方向线(pdm、mdm)
收盘价指标	收盘价移动平均(close_-1_s、close_-1_d、close_20_sma、close_20_mstd、close_10_sma、close_50_sma、close_2_sma)
K 线指标	开盘价(open)、最高价(high)、最低价(low)、成交量(volume)
重叠研究	移动平均价(SMA、MA2、WMA)、指数平均数(EMA、DEMA)、布林曲线(BOLL)

由于输入特征较多，而各个变量间存在信息冗余，从而影响 LSTM 的预测准确率及收敛速度。因此首先计算指标与收盘价的皮尔逊相关系数，保留相关系数较高的前 40 个指标。由于皮尔逊相关系数检验

的是线性相关性，而指标对于收盘价的影响大多都是非线性的。Reshaef 等人[4]在 2011 年提出的 MIC 可以捕捉每个特征与目标变量之间的任意关系，比如线性和非线性关系、函数或非函数关系等。两变量的 MIC 越大，则相关性越强，当两个变量具有严格确定的关系即 $y=f(x)$ 时，MIC 取 1，当两变量独立时，MIC 为 0。计算剩余的 40 个指标与收盘价的最大互信息系数，得到 MIC 前 20 的指标如下表 2：

Table 2. MIC top 20 indicators
表 2. MIC 前 20 的指标

close_2_sma	0.938	close_20_sma	0.798
high	0.931	boll	0.797
low	0.920	OBV	0.757
open	0.907	close_50_sma	0.687
SMA	0.904	tr	0.408
close_-1_s	0.897	volume	0.398
EMA	0.874	atr	0.253
DEMA	0.858	close_20_mstd	0.186
close_10_sma	0.857	ROC	0.177
middle_14_sma	0.829	dma	0.142

初步特征筛选后，通过随机森林算法构建股票收盘价预测模型，用于确定辅助变量最优特征子集。随机森林最早由 Breiman 提出[5]，用它计算特征重要性的思想是取每个特征在随机森林中的每颗树上做的平均贡献，最后比较特征的贡献大小，贡献通常用袋外数据错误率来衡量[6]，其基本步骤如下：

1) 从 N 个原始训练集中用 Bootstrap 有放回抽 n 个样本，进行 k 次采样，生成 k 个训练子集 D_k。

2) 从原始特征中随机抽取 m 个特征，对 D_k 进行训练，将 m 个特征作最优切分得到 K 棵决策树预测结果。

3) 计算特征重要性并按降序排序：重复抽样得到的一组数据来训练决策树，剩余数据计算第 i 棵决策树袋外错误样本数 (ErrOOB_i)。在保持其他特征不变的同时，对 OOB 中的 X^j 加入噪声干扰得到 OOB_i^j，再次计算袋外数据误差 $\overline{\mathrm{ErrOOB}_i^j}$。重复上述步骤，得到 $\overline{\mathrm{ErrOOB}_i^j}$ 与 $\left\{\mathrm{ErrOOB}_i^j \mid i=1,2,K\right\}$。计算所有决策树特征 X^j 置换前后 OOB 分类误差率的平均变化量：$\mathrm{VI}\left(X^j\right)=\frac{1}{K}\sum_i\left(\overline{\mathrm{ErrOOB}_i^j}-\mathrm{ErrOOB}_i^j\right)$。

4) 根据重要性剔除一定数量特征，剩余作为新的特征集，用此特征集重复上述过程，得到对应的特征集和袋外误差，直到剩下 m 个特征，选择误差最低的特征集作为 LSTM 的输入变量。

2.2. 股票走势预测

将多步特征选择的最优特征集作为股票预测模型的输入，由于股票时间序列前面的价格会影响后期价格走势，因此选择具有记忆性的 LSTM 神经网络，它是 Hochreiter 在 1997 年对循环神经网络的改良[7]，通过反向传播算法进行训练，并使用记忆单元来解决梯度消失的问题。其神经元结构[8]如图 2。

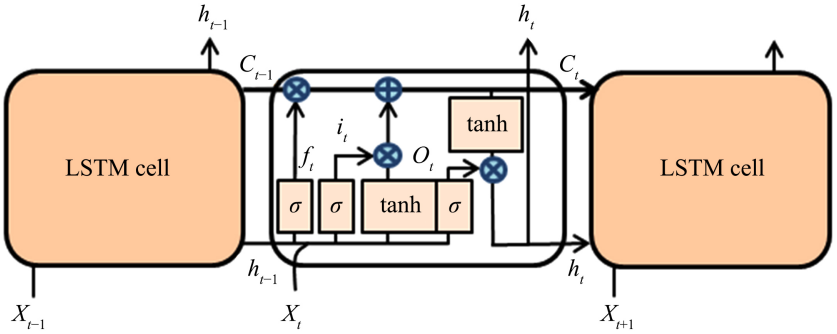

Figure 2. LSTM neuron structure
图 2. LSTM 神经元结构图

细胞状态C_t用于保存当前信息，并在下一时刻传递给 LSTM，通过门控单元传递信息。这些门决定了对过去和新信息的记忆遗忘程度，使 LSTM 成为一个长期依赖函数。遗忘门$\left(f_t\right)$控制上一时刻传递到当前时刻中的信息，输出为$f_t * C_{t-1}$

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

输入门$\left(i_t\right)$控制当前输入新信息$\overline{C_t}$中有多少可以加入到细胞状态中。Tanh 层用来产生当前的新信息，sigmoid 层用来控制有多少新信息可以传递给细胞状态。输出为$i_t * \overline{C_t}$

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

$$\overline{C_t} = \tanh\left(W_c\left[h_{t-1}, x_t\right] + b_c\right) \tag{3}$$

更新后的细胞状态来自上一时刻旧的细胞状态信息C_{t-1}和当前输入新信息：

$$C_t = f_t * C_{t-1} + i_t * \overline{C_t} \tag{4}$$

输出门$\left(O_t\right)$：更新细胞状态后，输出隐藏状态h_t，用 sigmoid 层来控制细胞状态信息，将细胞状态缩放至(−1, 1)作为隐藏状态的输出

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

$$h_t = O_t * \tanh\left(C_t\right) \tag{6}$$

为了使预测结果更加准确，采用遗传算法优化 LSTM 参数[9]。遗传算法是元启发式随机优化方法，基本思想是模拟生态的进化过程，通过交叉、变异等手段，有效地获得网络参数的近似全局最优解[10]。优化的网络参数有：LSTM 的层数，隐藏层神经元个数，dense 层层数。其初始参数设置为：交叉概率 0.5，变异概率 0.01，种群大小 20，世代大小 30，每条染色体长度为 8，染色体个数为 2。适应度为均方误差的倒数：

$$E_p = \frac{1}{\mathrm{MSE}} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{7}$$

MSE 表达实际值与期望值之间的误差，均方误差倒数越大表明网络性能越好，剔除低适应度低的个体，繁殖高适应度的个体。算法结束后，适应度值最大为最优解，即 LSTM 的最优参数。遗传算法优化 LSTM 结构如图 3。

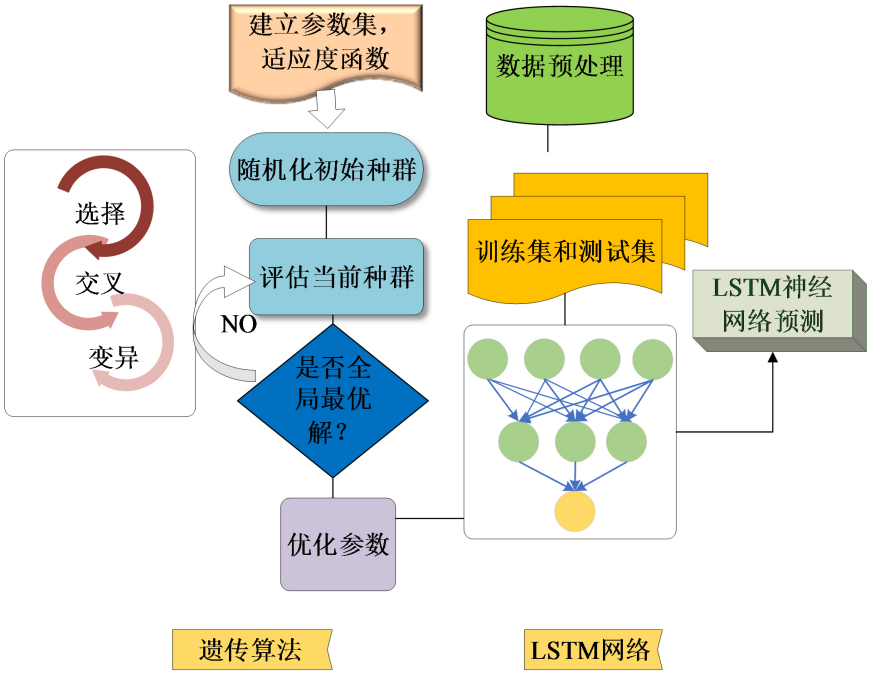

Figure 3. The flow chart of GA-LSTM
图 3. 混合 GA-LSTM 预测模型的框架

3. 数据与评价指标

3.1. 数据选择与处理

本文采集是不同行业代表企业十年的日数据，所有股票数据来自开源 python 金融数据接口 Tushare 包，主要实现对股票等金融数据采集、清洗、处理和储存。股票的所有技术指标都来自一个 Python 金融量化的高级库，名为 TA-Lib 的技术分析库。对收集到的数据作如下处理：

1) 数据清洗与拆分

本次研究中使用的数据是通过 API 和网站收集的，所以有些值缺失或没有意义。因此，在适当的时候，将计算出的统计平均值作为观测值替换缺失的数据，为保证数据真实性，不对周末数据作缺失处理。在训练过程中，将数据的 70%数据作为训练集来估计模型，剩下的 30%作为测试集来测试最优模型的性能。

2) 数据转换

数据归一化是数据预处理的重要步骤。归一化的目的将数据压缩到(0, 1)范围内。为了更好的说明问题，我们用式 8 对数据进行归一化，目的是统一量纲，方便计算，减少梯度和加速收敛，其中 X_i 代表第 i 天的收盘价，$X_{\max}$ 和 $X_{\min}$ 代表 X_i 的最大值和最小值。

$$\hat{X}_i = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}},\ i = 1, 2, \cdots, N \tag{8}$$

在建立 LSTM 模型的过程中，归一化数据需要进行滑动窗口处理，每组 X 对应一个 Y，假设原始时间序列 $X = \{x_1, x_2, x_3, \cdots, x_n\}$ 的宽度时间窗 l，预测周期为 pr，数据会生成 nl-pr + 1 个时间窗序列，分别为 $\{x_1, \cdots, x_l\}, \{x_2, \cdots, x_{l+1}\}, \cdots, \{x_{n-k-j+1}, \cdots, x_n\}$。在本文中，最合适的时间窗口宽度初始设置为 16，如图 4 所示，每组监督数据包括一组 $\{x\}$ 和一组 y，每组 $\{x\}$ 包括一个从 i 到 $i+15$ 共 16 个指标数据，每个对应的 y 是 $i+16$ 个收盘价。构建的数据集作为 LSTM 模型的输入序列进行训练和验证。

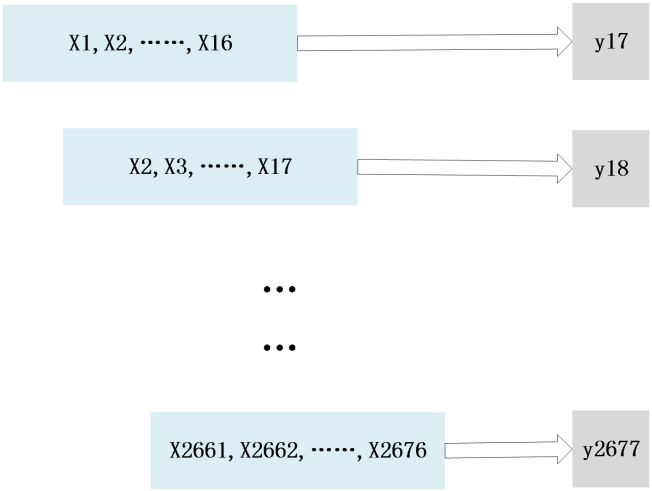

Figure 4. The construction of data set
图 4. 构建数据集

3.2. 预测结果评估

为了有效评价所提模型，用四个指标对预测结果评估。均方根误差：作为股价预测模型的评价标准，同时作为训练模型时的损失函数，RMSE 的值越小，模型的拟合效果越好。平均绝对误差百分比：MAPE 越低，模型的预测结果越可靠。决定系数：R^2 越接近于 1，模型拟合效果越好，R^2 越接近于 0，模型拟合效果越差，当 R^2 为 1 时，样本数据完全拟合。y_i 是序列实际值，$\hat{y}_i$ 是其预测值，$\bar{y}_i$ 是原始序列平均值，指标计算如下：

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2} \tag{9}$$

$$\mathrm{MAPE}=\frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_i-\hat{y}_i\right|}{y_i}\times 100 \tag{10}$$

$$R^2=1-\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}_i\right)^2} \tag{11}$$

4. 实验设计与结果分析

本节主要分为三个部分。第一部分以中国银行股票数据为例，展示所提模型预测结果。第二部分将所提出的多步特征选择方法与其他降维方法筛选出的变量作为输入变量，再导入 LSTM 网络进行预测方法进行对比，第三部分利用 ARIMA、SVR、RNN、GRU、MLP 和 LSTM 四个模型对股票收盘价进行预测。第四部分选择中国股市不同行业的股票数据运进行预测从而验证此模型的普适性。结果表明，使用多步特征选择降维后的 LSTM 网络具有更好的预测效果。

4.1. 特征选择结果

通过皮尔逊相关系数和互信息系数初步筛选特征后用随机森林计算特征重要性。随机森林参数用二次划分贪心选择算法优化。首先，经一次划分随机搜索算法，计算在不同的决策树个数进和最大深度下的 MSE，MAE。确定决策树个数在 300~400，最大深度为 10~120。图 5 中标记的平均测试分数的最高点也是平均绝对误差最低值 MAE = 0.0066，对应的每个划分最少的样本数为 7，叶子节点最少的样本数为

6。同理缩小参数取值范围进行二次网格搜索，计算每个网格的均方根误差，得到最优决策树个数为 400，决策树最大深度和单棵决策树最大特征数分别为 140 和 34。

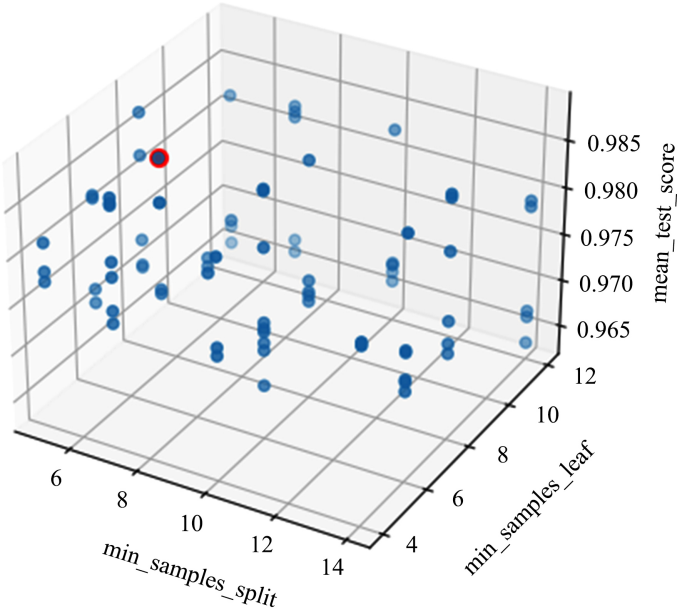

Figure 5. The scope of min_samples_split, min_samples_leaf value
图 5. 每个划分最少的样本数、叶子节点最少的样本数取值

使用优化后的随机森林计算特征的重要性如图 6 所示。确定 LSTM 模型的最终输入变量为 close_2_sma、high、low 和 SMA。

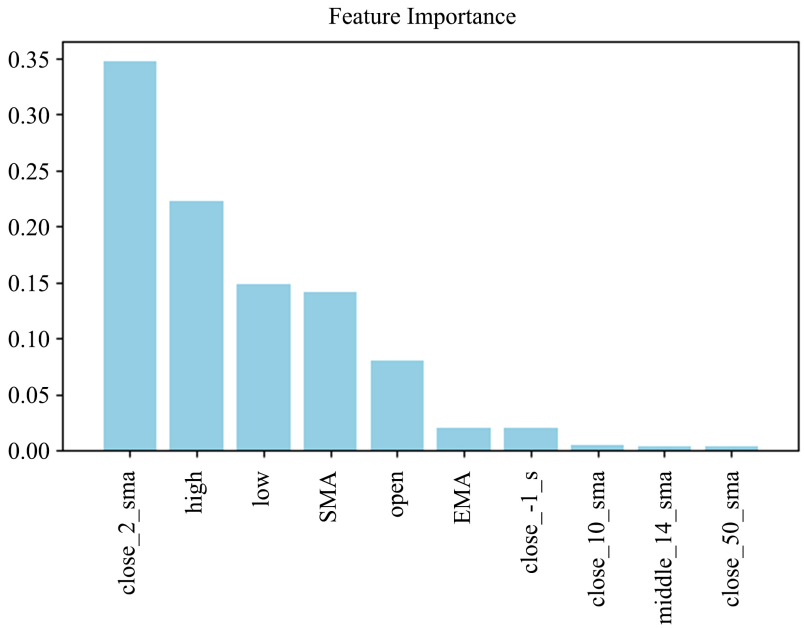

Figure 6. The importance of various features
图 6. RF 特征选择特征重要性

4.2. GA-LSTM 预测

以中国银行股票的收盘价预测为例，输入变量包括了上一步特征选择得到的四个重要变量。利用遗传算法优化 LSTM 模型得到模型参数为：LSTM 的层数为 3，隐藏层的神经元个数为 138，dense 层层数为 1，神经元个数为 43。对于其他参数：早停法获得迭代次数为 25，mini-batch 得到一次训练选择样本数为 16。使用 Adam 算法作为模型权重优化器，学习率为 0.0001。用上述参数训练模型得到最终拟合结果如图 7。

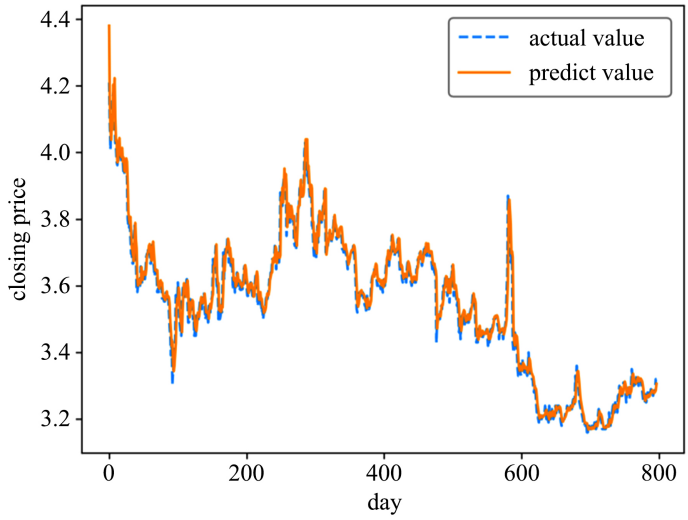

Figure 7. Bank of China stock forecast results
图 7. 中国银行股票预测结果

由图 7 可以看出两条曲线几乎重合，初步认为此模型的预测精度较高。图 8(a)损失函数图看出，训练集和测试集的损失函数都已经收敛且很接近，说明模型能够很好地拟合训练样本和测试样本的分布，泛化能力强。从图 8(b)预测的残差图来看，残差值绝对值分布在 0.1 以内，个别位置有较大波动，误差较小且较稳定，进一步说明模型的优越性。

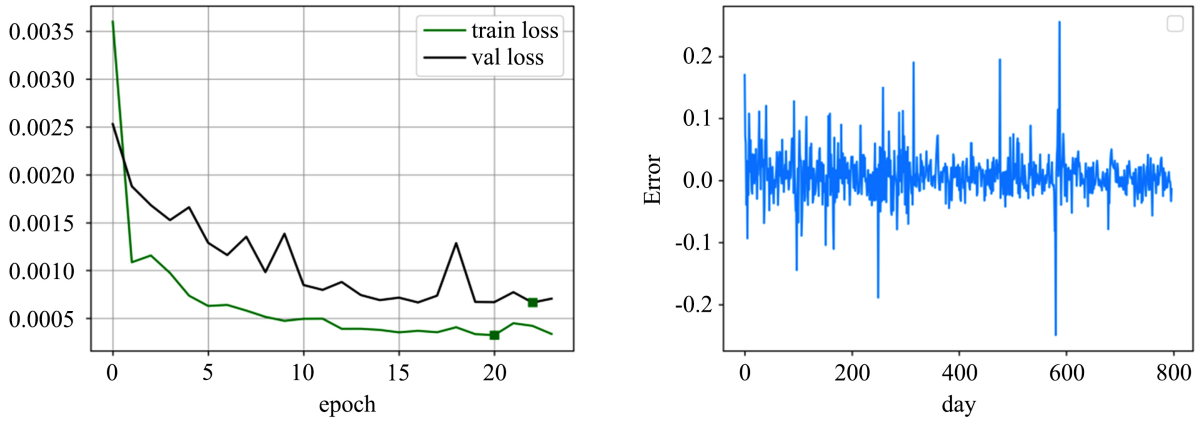

Figure 8. Loss function graph and error analysis
图 8. 损失函数和误差分析

4.3. 降维模型对比

将多步特征选择与主成分、LASSO 降维后的输入变量导入 LSTM 神经网络，使用共同的模型参数，预测结果及部分序列放大图的预测值和实际值如图 9 所示，可以看出与真实序列相比，所提出的模型，预测曲线(黑色)与实际历史曲线(红色)上下交织，没有明显的相位差，相较于其他曲线与实际值更加接近。表明多步特征选择降维再进行预测有更好的性能。

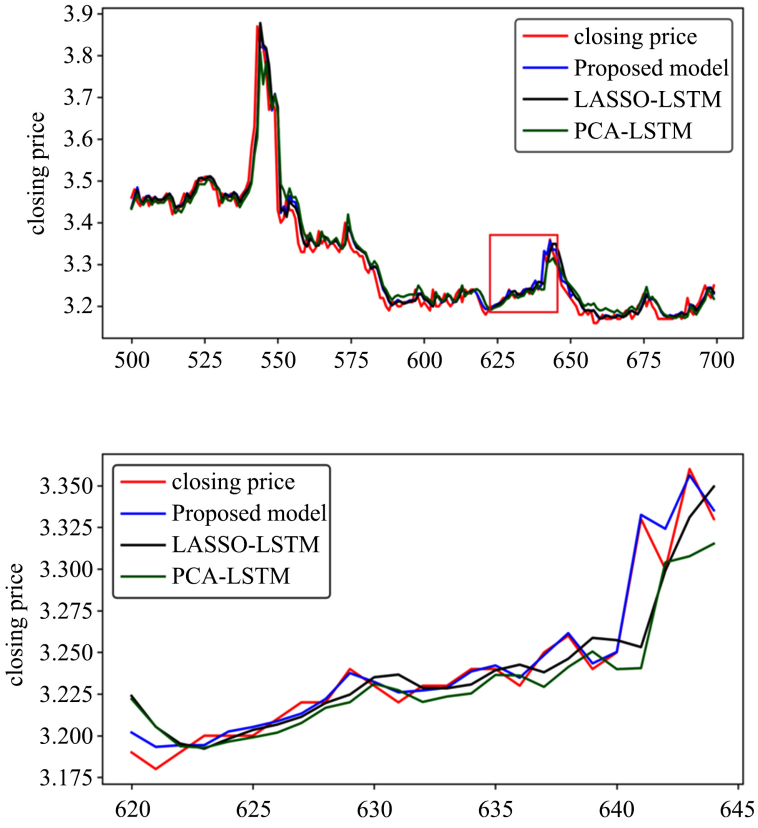

Figure 9. Comparison of dimensionality reduction model
图 9. 降维模型预测对比图

为了定量评估预测精度，列出了这些模型的三个评估指标见表 3。可以发现，所提出的多步特征选择相较于其他模型 RMSE、MAPE 都较小，R^2 说明拟合优度较好，其次是 LASSO 降维的方法。

Table 3. Comparison of dimensionality reduction model
表 3. 中国银行降维模型预测对比

model	R^2	test		train	
		RMSE	MAPE	RMSE	MAPE
Proposed model	0.998	0.014	0.49	0.021	0.34
PCA-LSTM	0.994	0.036	1.09	0.051	0.72
Lasso-LSTM	0.968	0.035	1.06	0.053	0.65

4.4. 预测模型对比

将本文所提的 GA-LSTM 模型与自回归移动平均(ARIMR)模型、多层感知机(MLP)模型、循环神经网络(RNN)模型、支持向量回归(SVR)模型和门控单元网络(GRU)的评测结果相对比，分析了所提模型的性能。模型的输入变量都为特征选择预处理后的变量。预测对比图及细节放大图如图 10 所示。与真实序列相比，各种深度学习模型的预测的结果表现为右移，说明预测较实际有一定时间的滞后性，其中最为明显的是线性模型 ARIMA。对于所提出的模型，预测曲线(黑色)与实际历史曲线(红色)上下交织，没有明显的相位差，表明预测效果较好。

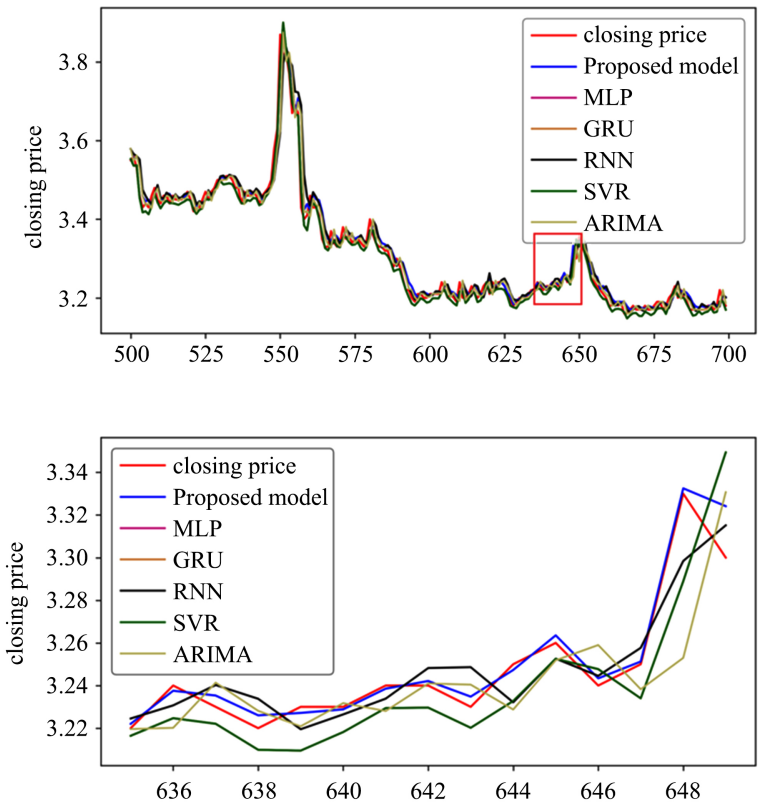

Figure 10. Comparison of predictions of parallel models
图 10. 并行模型预测对比图

为了定量评估预测精度，表 4 列出了这些模型的三个性能标准。可以发现，传统模型的 RNN 和 ARIMR 的预测偏差高于其他模型，而基于 SVR 和 GRU 的预测偏差高于所提模型。该模型 MAPE、RMSE 和 R^2 分别为 0.34、0.0145 和 0.9986。所提模型预测效果优于 ARIMA 模型和 RNN 模型，相较于深度学习模型中的 MLP、SVR、GRU 预测效果有所提高。

4.5. 模型普适性检验

为了测试交易算法在中国股市中的表现，我们选择中国十大不同行业具有代表性公司的股票价格进行了实验。这些股票具有很强的流动性，能反映该行业的发展方向和整体走向，为交易策略提供良好的目标。它们代表了中国股市的发展方向，对投资者具有吸引力，预测结果如表 5。可以看出，各行业数据计算预测误差小，拟合优度高。因此，该模型应用于不同的数据时，可以取得更好的预测结果。

Table 4. Comparison of parallel model
表 4. 中国银行并行模型预测对比

model	R^2	test		train	
		RMSE	MAPE	RMSE	MAPE
Proposed model	0.998	0.014	0.34	0.021	0.49
ARIMA	0.986	0.032	0.55		
MLP	0.989	0.019	0.37	0.029	0.63
RNN	0.983	0.036	0.69	0.053	0.99
SVR	0.991	0.018	0.42	0.023	0.51
GRU	0.987	0.022	0.41	0.037	0.72

Table 5. Comparison of rolling forecast results of stock prices
表 5. 股票价格滚动预测结果对比

行业	企业	样本	MAPE	RMSE	R^2
金融	中国银行	2677	0.008	0.015	0.984
房地产	万科 A	2628	0.026	1.511	0.969
教育	中公教育	2557	0.048	0.434	0.953
建筑	中国铁建	2680	0.031	0.419	0.972
交通运输	春秋航空	1470	0.018	1.224	0.984
农业	牧原乳业	2565	0.022	0.836	0.986
餐饮食宿	锦江旅馆	2611	0.021	0.525	0.969
制造业	中兴通讯	2614	0.035	0.774	0.968
信息技术	科大讯飞	1048	0.025	1.453	0.671
农林牧渔	万达影业	1708	0.025	0.576	0.986

5. 结语

本文针对非平稳、噪声较大的金融时间序列数据，提出了一种混合机器学习多步特征选择的GA-LSTM 框架，用于多指标下股票收盘价的预测，以提高预测精度。实验选取中国股市各行业数据，通过数据收集与处理、特征选择、网络模型参数优化、网络模型建立和结果评价分析五个过程验证所提模型的预测效果。结果表明，与 PCA-LSTM、LASSO-LSTM 降维方法相比，本文通过选取大量技术指标，首先利用多步特征选择大大减少模型的输入维数，计算特征重要性，提高了预测准确率；与机器学习模型 MLP、ARIMA、RNN、SVR、GRU 相比，提高了预测精度。未来工作可以尝试以下两个方面，一是选取更多种类的算法探讨预测模型的性能以及挖掘重要性预测指标；二是将模型应用到高频数据或不同周期的数据中。

参考文献

[1] Wang, J. and Wang, J. (2015) Forecasting Stock Market Indexes Using Principle Component Analysis and Stochastic Time Effective Neural Networks. *Neurocomputing*, **156**, 68-78. https://doi.org/10.1016/j.neucom.2014.12.084

[2] Cavalcante, R.C., Brasileiro, R.C., Souza, V.L.F., Nobrega, J.P., and Oliveira, A.L.I. (2016) Computational Intelligence and Financial Markets: A Survey and Future Directions. *Expert Systems with Applications*, **55**, 194-211. https://doi.org/10.1016/j.eswa.2016.02.006

[3] Niu, T., Wang, J., Lu, H., Yang, W. and Du, P. (2020) Developing a Deep Learning Framework with Two-Stage Feature Selection for Multivariate Financial Time Series Forecasting. *Expert Systems with Applications*, **148**, Article ID: 113237. https://doi.org/10.1016/j.eswa.2020.113237

[4] Kinney, J.B. and Atwal, G.S. (2014) Equitability, Mutual Information, and the Maximal Information Coefficient. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 3354-3359. https://doi.org/10.1073/pnas.1309933111

[5] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324

[6] Nguyen, C., Wang, Y. and Nguyen, H.N. (2013) Random Forest Classifier Combined with Feature Selection for Breast Cancer Diagnosis and Prognostic. *Journal of Biomedical Science and Engineering*, **6**, 551-560. https://doi.org/10.4236/jbise.2013.65070

[7] Chen, K., Zhou, Y. and Dai, F. (2015) A LSTM-Based Method for Stock Returns Prediction: A Case Study of China Stock Market. 2015 *IEEE International Conference on Big Data* (*Big Data*), Santa Clara, CA, 29 October-1 November 2015, 2823-2824. https://doi.org/10.1109/BigData.2015.7364089

[8] Van Houdt, G., Mosquera, C. and Nápoles, G. (2020) A Review on the Long Short-Term Memory Model. *Artificial Intelligence Review*, **53**, 5929-5955. https://doi.org/10.1007/s10462-020-09838-1

[9] Chung, H. and Shin, K.S. (2018) Genetic Algorithm-Optimized Long Short-Term Memory Network for Stock Market Prediction. *Sustainability*, **10**, 3765. https://doi.org/10.3390/su10103765

[10] Sharma, D.K., Hota, H.S., Brown, K. and Handa, R. (2021) Integration of Genetic Algorithm with Artificial Neural Network for Stock Market Forecasting. *International Journal of System Assurance Engineering and Management*. https://doi.org/10.1007/s13198-021-01209-5