基于 BO-CNN-LSTM 的股价预测研究

万梦洁

【摘要】为提高股价预测的精确度，本文提出贝叶斯优化的卷积神经网络 - 长短期记忆神经网络股价预测模型。基于卷积神经网络（Convolutional Neural Network，CNN）、长短期记忆神经网络（Long Short-Term Memory，LSTM）的原理特性，将二者串行结合，构建 CNN-LSTM 模型。针对 CNN-LSTM 模型参数多，人工调优难的问题，利用贝叶斯优化（Bayesian Optimization，BO）寻找最优超参数组合，构建 BO-CNN-LSTM 模型。实验表明经过贝叶斯优化后的 CNN-LSTM 模型 R^2 提高了 1.923%，对股价预测更优。

【关键词】股价预测；贝叶斯优化；卷积神经网络；长短期记忆神经网络

股票数据呈现为时间序列形式，差分移动平均模型（Autoregressive Integrated Moving Average，ARIMA）、广义自回归条件异方差模型 (Generalized Autoregressive Conditional Heteroskedasticity，GARCH) 等可以对其进行预测。杨琦、曹显兵利用 ARMA-GARCH 模型预测大众公用股票，发现 ARIMA 与 GARCH 模型的结合可以提高预测准确性。然而时间序列通常是非线性、非平稳的，传统模型在时序预测上存在局限性。

近年来，机器学习在股票预测领域得到了广泛应用。韩旭等结合商空间和支持向量机（Support Vector Machines，SVM）预测黄金价格，该方法比传统模型预测效果更好。随着深入研究，深度学习成为目前最先进的技术之一。陈祥一通过 CNN 预测沪深 300 指数的涨跌，相较传统机器学习表现更佳。彭燕等利用 LSTM 模型预测苹果公司的股价，提升了预测的准确率。然而单一的神经网络被证实存在一定的局限性，无法兼备提取时序数据的长期依赖性和关键特征的能力。混合网络模型可以有效整合单个模型的优点，从而提升模型的鲁棒性和预测能力。基于此，本文以沪深 300 指数为例，结合 CNN、LSTM 在空间、时间维度上的特征提取能力，同时针对调参难的问题，使用贝叶斯算法优化 CNN-LSTM 模型的超参数，构建 BO-CNN-LSTM 模型。

一、基本原理

1. 卷积神经网络（Convolutional Neural Network，CNN）

CNN 主要由卷积层、池化层和全连接层组成。卷积层为其核心，卷积层利用卷积核对输入特征进行卷积操作，实现特征提取。

2. 长短期记忆神经网络（Long Short-Term Memory，LSTM）

LSTM 主要包含记忆细胞和输入门、遗忘门、输出门三个门控结构。

遗忘门：选择性地遗失不需要的信息。

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

作者简介：万梦洁（2001—），女，重庆万州人，汉族，硕士研究生，成都信息工程大学统计学院，研究方向：数据挖掘

输入门：决定输入层信息是否进入记忆细胞。

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

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (3)$$

细胞状态：确定新信息的更新。

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (4)$$

输出门：决定输出信息。

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t * \tanh(C_t) \quad (6)$$

x_t 为第 t 时刻输入数据，h_{t-1} 为 t-1 时刻隐藏层状态，f_t 为 t 时刻遗忘门控制值，i_t 为 t 时刻输入门控制值，C_t 为 t 时刻细胞状态候选值，C_{t-1} 为 t-1 时刻细胞状态，C_t 为 t 时刻细胞状态实际值，o_t 为 t 时刻输出门控制值，h_t 为 t 时刻隐藏状态输出向量，W_f、W_i、W_C、W_o 为权重矩阵，b_f、b_i、b_C、b_o 为偏置向量，σ 为 sigmoid 函数，tanh 为 tanh 函数。

3. 贝叶斯优化（Bayesian Optimization，BO）

贝叶斯优化算法优化超参数，避免了人工调参带来的局限性。给定优化的目标函数，不断添加样本点更新目标函数的后验分布。以 f(x) 作为超参数的目标函数，$X=x_1,x_2,x_3,\cdots\cdots,x_n$ 为一组超参数组合。找到 $x \in X$ ，使得:

$$x^* = argminf(x) \quad (7)$$

x^*为最优超参数集。

二、实证分析

本文数据来自英为财情（Investing.com），选取深沪 300 指数（SH000300）2010 年 1 月 4 日到 2024 年 1 月 25 日，共 3421 的交易日数据，获取特征有开盘价、最高价、最低价、收盘价、成交量、涨跌幅。构造的输入为预测日前 5 天影响收盘价的变量数据，输出为预测日收盘价。将 2010 年 1 月 4 日到 2020 年 7 月 22 日 2566 条数据作为训练集；2020 年 7 月 23 日到 2022 年 8 月 30 日 513 条数据作为验证集；2022 年 8 月 31 日到 2024 年 1 月 25 日 342 条数据作为测试集。

1. 特征选取

本文旨在利用股票历史数据来预测未来股票收盘价，通过对数据特征进行相关性分析，可以观察到收盘价与开盘价、最高价、最低价、成交量之间相关性较高，与涨跌幅的相关性较低，因此最终选择的特征为开盘价、最高价、最低价、收盘价、成交量。相关系数矩阵如表 1 所示。

表 1 相关系数矩阵

	开盘价	最高价	最低价	收盘价	成交量	涨跌幅
开盘价	1.000000	0.999389	0.999221	0.998519	0.523473	−0.017045
最高价	0.999389	1.000000	0.998928	0.999275	0.535508	0.003416
最低价	0.999221	0.998928	1.000000	0.999245	0.512863	0.006364
收盘价	0.998519	0.999275	0.999245	1.000000	0.526449	0.029986
成交量	0.523473	0.535508	0.512863	0.526449	1.000000	0.073396
涨跌幅	−0.017045	0.003416	0.006364	0.029986	0.073396	1.000000

2. 数据归一化

本文使用 Min-Max 归一化方法消除变量量纲的影响。

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (8)$$

x为原始数据，x_{min}、x_{max}分别为最大值、最小值。

3. 模型构建

（1）CNN-LSTM 模型

将卷积神经网络与长短期记忆神经网络结合，对股票收盘价做短期预测。首先通过卷积层和最大池化层提取输入数据的特征，将其输入 LSTM 网络层，其次经过第一个全连接层（Dense）以及防止模型过拟合而增加的 Dropout 层，最后通过全连接层（Dense）输出预测结果。CNN-LSTM 模型结构如图 1 所示。

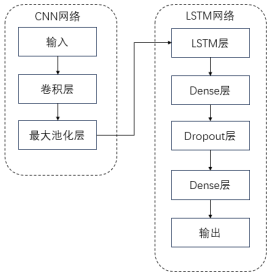

图 1 CNN-LSTM 模型结构

（2）BO-CNN-LSTM 模型

为了解决 CNN-LSTM 模型中参数众多，人工难以调优的问题，本文构建了 BO-CNN-LSTM 模型。首先对股票数据进行预处理，将数据划分为训练集、验证集、测试集；其次将训练集输入 CNN-LSTM 网络中进行训练，验证集用于在训练过程中验证模型性能，最小化损失函数，同时使用贝叶斯优化算法调整 CNN-LSTM 网络的超参数，得到最优的超参数组合；最后使用最优化超参数组合的模型对测试集进行测试。BO-CNN-LSTM 组合模型的预测流程如图 2 所示。

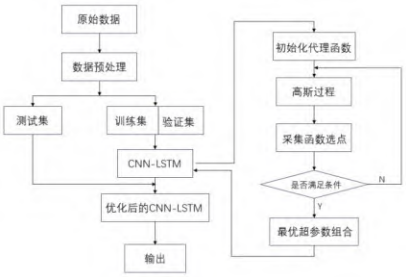

图 2 基于 BO-CNN-LSTM 组合模型的预测流程

利用贝叶斯算法，通过计算返回目标函数中的最小值对 CNN-LSTM 模型中的学习率、卷积核个数、卷积核尺寸大小、LSTM 层神经元个数、第一个 Dense 层神经元个数、批大小六个超参数进行组合优化。得到目标函数最小值下的最优超参数组合如表 2 所示。

表 2 贝叶斯优化参数组合

超参数	设定范围	值
卷积核大小	[1，5]	2
卷积核数量	[8，64]	22
LSTM 层神经元个数	[32，128]	33
Dense 层神经元个数	[32，128]	123
批大小	[32，256]	70
学习率	[0.001，0.01]	0.004

由表 2 可以看出，基于贝叶斯优化之后的最优超参数组合卷积核大小、卷积核数量、LSTM 层神经元个数、Dense 层神经元个数、批大小、学习率分别为 2、22、33、123、70、0.004。经过多次实验将迭代次数 (epoch) 设置为 41，确保实验误差收敛。

4. 评价指标

本文选取平均绝对误差 (MAE)、均方误差 (MSE)、均方根误差 (RMSE)、决定系数 (R^2) 评估模型。

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{9}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{10}$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{11}$$

$$R^2 = 1 - \frac{\Sigma(\hat{y}_i - y_i)^2}{\Sigma(\hat{y}_i - \bar{y})^2} \tag{12}$$

n为样本个数，y_i为第 i 个样本的真实值，$\hat{y}_i$为第 i 个样本的预测值，$\bar{y}$为样本平均值。

5. 实验结果

用测试集对 BO-CNN-LSTM 模型进行测试，将该模型与 CNN-LSTM、CNN、LSTM 的预测结果对比。

由图 3 可见，BO-CNN-LSTM 模型的拟合曲线更接近真实值，拟合效果更好。相比之下，CNN、LSTM、CNN-LSTM 这几种模型的拟合曲线与真实值之间存在的偏差更大。因此，本文构建的 BO-CNN-LSTM 模型的预测效果更好。

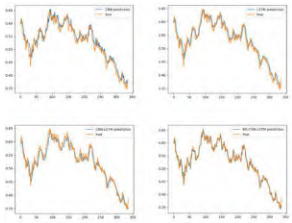

图 3 不同模型预测结果对比

表 3 模型评价指标对比

模型	R^2	MAE	MSE	RMSE
CNN	0.94364	0.01363	0.00030	0.01718
LSTM	0.95186	0.01289	0.00025	0.01588
CNN-LSTM	0.95624	0.01186	0.00023	0.01514
BO-CNN-LSTM	0.97547	0.00854	0.00013	0.01133

由表 3 可见，BO-CNN-LSTM 模型的 MAE、MSE、RMSE 值分别为 0.00854、0.00013、0.01133，其值均小于其他三种模型。此外，BO-CNN-LSTM 模型的 R^2 达到了 0.97547，相较于其他三种模型分别提高了 3.183%、2.361%、1.923%，进一步印证了本文构建的 BO-CNN-LSTM 模型在股价短期预测方面有着更高的精度。

三、结论

本文提出基于贝叶斯优化的 CNN-LSTM 模型，利用卷积、池化等操作提取数据特征，再将其输入到 LSTM 中，同时用贝叶斯算法寻找模型的最优超参数组合，最终构建了 BO-CNN-LSTM 模型，对沪深 300 指数进行预测。实验对比了 BO-CNN-LSTM 模型与其他三种模型的预测效果，结果表明，BO-CNN-LSTM 模型比其他三种模型的预测效果更优，相比 CNN-LSTM 模型 MAE、MSE、RMSE 值分别下降 0.332%、0.01%、0.381%，R^2 提高了 1.923%。利用该模型能够有效提高股价预测的精度。

参考文献

[1] 冯盼，曹显兵 . 基于 ARMA 模型的股价分析与预测的实证研究 [J]. 数学的实践与认识，2011,41(22):84-90.

[2] 韩旭，杨珊，王喜梅 . 基于商空间的黄金价格 SVM 模型预测 [J]. 黄金科学技术，2020,28(1):148-157.

[3] 陈祥一 . 基于卷积神经网络的沪深 300 指数预测 [D]. 北京：北京邮电大学，2018.

[4] 彭燕，刘宇红，张荣芬 . 基于 LSTM 的股票价格预测建模与分析 [J]. 计算机工程与应用，2019,55(11):209-212.

[5] 景楠，史紫荆，舒毓民 . 基于注意力机制和 CNN-LSTM 模型的沪铜期货高频价格预测 [J/OL]. 中国管理科学：1-13[2023-03-12].