E-Commerce Letters 电子商务评论, 2024, 13(2), 2253-2258
Published Online May 2024 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2024.132275

Hans汉斯

基于机器学习的股票价格预测研究

贾雨菲

贵州大学经济学院，贵州 贵阳

收稿日期：2024年3月18日；录用日期：2024年4月8日；发布日期：2024年5月24日

摘　要

股票价格预测可以提供有关未来市场走势的信息，对投资者而言具有十分重要的影响，将机器学习引入股票价格预测中，有助于投资者制定更明智的投资决策。基于此，本文随机选取了一只股票——中国银行在2018年11月16日~2023年11月15日期间的数据作为样本，通过ARIMA模型、支持向量机(SVM)模型、LSTM模型对其股价走势进行预测，最后得出基于LSTM模型的深度神经网络模型具有较好的预测精度。

关键词

股票价格预测，ARIMA，LSTM，支持向量机

Research on Stock Price Prediction Based on Machine Learning

Yufei Jia

School of Economics, Guizhou University, Guiyang Guizhou

Received: Mar. 18th, 2024; accepted: Apr. 8th, 2024; published: May 24th, 2024

Abstract

Stock price prediction can provide information about the future market trend, which has a very important impact on investors. Introducing machine learning into stock price prediction can help investors make more intelligent investment decisions. Based on this, this paper randomly selects the data of a stock—Bank of China from November 16, 2018 to November 15, 2023 as a sample, and predicts its stock price trend through ARIMA model, support vector machine (SVM) model and LSTM model. Finally, it is concluded that the deep neural network model based on LSTM model has better prediction accuracy.

文章引用: 贾雨菲. 基于机器学习的股票价格预测研究[J]. 电子商务评论, 2024, 13(2): 2253-2258.
DOI: 10.12677/ecl.2024.132275

Keywords

Stock Price Prediction, ARIMA, LSTM, Support Vector Machine

1. 引言

随着金融市场的不断发展和信息技术的快速进步，股票价格的预测成为投资者和交易者关注的焦点之一。股票市场的动态和复杂性使得准确预测股票价格变得极具挑战性。传统的金融模型和方法在面对大规模、高维度的市场数据时往往显得力不从心，因而寻找更为有效的预测方法成为当务之急。

近年来，机器学习技术以其出色的数据处理和模式识别能力引起了广泛关注，成为股票价格预测研究的一项重要工具。机器学习不仅能够处理大量历史数据，还可以自动学习和适应市场的变化，为投资者提供更全面、精准的市场分析。基于此，本文随机选取了一只股票——中国银行 2018 年 11 月 16 日到 2023 年 11 月 15 日共计 1213 个交易日的收盘价作为样本数据，采用 ARIMA 模型、支持向量机(SVM)模型和长短期记忆网络模型(LSTM)分别对其股票价格进行预测分析，最后通过对比选出股票价格预测的效果最好的模型。本文旨在探讨机器学习在股票价格预测中的应用，在为投资者、金融从业者和学术界提供有关机器学习在股票价格预测中应用的深入理解，并为未来的研究和实践提供有益的启示方面具有一定价值和意义。

2. 文献综述与模型介绍

(一) 文献综述

股票价格预测一直是金融和经济学领域的重要议题。作为典型的金融时间序列，股票数据多年来一直吸引学者们开发更精准预测模型的研究兴趣。股票行为的早期研究可以追溯到 1900 年的 Bachelier，他以随机游动的方式描述了股票价格的走势。Fama (1965)则通过检验股票价格的变化确认了其具有随机游动的特征[1]。传统的时间序列模型在预测时采用参数统计模型，例如自回归移动平均(ARMA)模型、差分自回归移动平均(ARIMA)模型和矢量自回归模型等[2]。这些模型旨在寻找最佳估计值。闫宇、吴海涛(2020)在使用纳斯达克综指数据时，首先对数据进行平稳处理，然后拟合了 ARIMA 模型，进行了对股票走势的短期预测[3]。尽管预测相对准确，但以上的 ARIMA 模型并未考虑其他因素对股票价格的影响，并且选择的是短期数据[3]。

而关于 SVM 模型预测方面，Vapnik (1968)首次提出了支持向量机的相关理论[4]。Muller (2003)的研究表明，支持向量回归机在预测方面表现优于其他时间序列方法[5]。Pai 和 Lin (2005)在股价预测中成功地将 ARIMA 模型和 SVM 模型融合，获得了显著的效果，比较显示复合模型的预测效果胜过单一模型[6]。阎威武和邵惠鹤(2003)在疾病诊断中应用了两种不同的 SVM，结果发现 SVM 模型表现出良好效果且具有潜力[7]。此外，顾红其(2010)和张玉(2011)分别利用 SVM 模型对我国期货价格和税收情况进行了预测[8] [9]。

在同一主题中，周凌寒(2018)的研究使用了基于历史交易数据的 LSTM 预测模型作为对照。研究结果显示，融合情感特征的 LSTM 模型相较于对照模型提高了 5%~6%的准确性，达到了 70.5%的水平[10]。

这项研究同时验证了 LSTM 在股票行情预测方面的有效性。另外，陆泽楠和商玉林(2017)利用了 10 年的钢铁交易价格数据，使用了 LSTM 和 SVR 两个模型进行预测，并从单因子和双因子两个角度进行了分析。结果显示，相对于 SVR 模型，LSTM 模型的预测耗时更短，精度更高。另外，双因子 LSTM 模型的预测误差也比单因子模型小[11]。总体而言，无论是国内还是国外，运用机器学习对股票价格进行预测的研究已有不少文献，但对于预测短期或长期时间序列的模型并没有统一的结论。

(二) 模型介绍

(1) ARIMA 模型

ARIMA (差分自回归移动平均)模型是一种经典的时间序列分析方法，用于预测未来的数据点。ARIMA 模型结合了自回归(AR)模型、差分(I, Integrated)运算和移动平均(MA)模型的特性。这个模型适用于一定程度上平稳的时间序列数据。

ARIMA 模型的三个主要部分分别是：自回归(AR)部分：表示当前观测值与先前观测值之间的关系。AR(p)模型包含 p 个滞后项，表示当前值与前 p 个时刻的值有关；差分(I, Integrated)部分：表示为使时间序列变得平稳而进行的差分操作。如果一次差分可以使数据平稳，那么模型中的差分阶数(d)为 1。如果需要多次差分才能使数据平稳，那么差分阶数将增加；移动平均(MA)部分：表示当前观测值与先前观测误差的线性组合。MA(q)模型包含 q 个滞后的误差项。ARIMA 模型的数学表达式如下：

$$X_t = c + U_t + \sum_{i=1}^{p} \alpha_i X_{t-i} - \sum_{i=1}^{q} \beta_i \varepsilon_{t-q} \tag{1}$$

ARIMA 模型的建模过程通常包括对时间序列数据进行可视化和平稳性检验，确定差分阶数，首先通过检查自相关和偏自相关图来选择 AR (自回归)和 MA (移动平均)模型的阶数。最后，拟合 ARIMA 模型并执行预测。

(2) SVM 模型

支持向量机(Support Vector Machine, SVM)是一种在监督学习中应用广泛的机器学习算法，它是一个强大的机器学习算法，适用于各种领域，包括图像分类、文本分类、生物信息学等，主要用于分类和回归分析。SVM 的核心理念是通过寻找一个超平面，将不同类别的样本有效地分隔，并确保最大程度地扩大两个类别样本点与超平面的距离。在训练过程中，给定一组标记的训练实例，每个实例被分为两个类别之一。SVM 训练算法生成一个非概率的二元线性分类器，该模型能够精确将新实例分类为其中一个类别。该模型将实例映射为空间中的点，以创建最大的明显间隔，有效地将同一类别的实例分隔开来。随后，新实例也被映射到同一空间，并根据其在间隔的哪一侧进行分类预测。为了处理原始特征空间中的线性不可分问题，SVM 利用核函数将数据映射到高维空间，常用的核函数包括线性核、多项式核以及径向基函数(RBF)核等。

(3) LSTM 模型

长短期记忆网络(Long Short-Term Memory, LSTM)是一种常用于处理序列数据的深度学习模型，特别适用于时间序列预测、自然语言处理等任务。LSTM 是循环神经网络(Recurrent Neural Network, RNN)的一种变体，设计用于解决 RNN 中的梯度消失和梯度爆炸等问题。LSTM 的设计重点在于能够捕捉和记忆长期依赖关系。它通过门控机制来控制信息的流动，使得模型能够更有效地保存和访问长期记忆。LSTM 包含三种门控单元：遗忘门(forget gate)、输入门(input gate)和输出门(output gate)。这些门控单元通过激活函数(如 sigmoid 函数)来控制信息的流动，从而实现对信息的筛选和控制。LSTM 中的记忆单元负责存储和更新信息。通过遗忘门来决定保留哪些信息，输入门来确定要添加哪些信息到记忆单元中，输出门则基于当前输入和记忆单元状态来生成输出。LSTM 模型是深度学习中的一种强大工具，能够有效地处理和建模序列数据中的复杂关系。常用于股票价格预测、天气预测、交通流量预测等时间序列数据的预

测任务。

遗忘门(forgetgate)的表达式为：

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{2}$$

f_t 为遗忘门的输出；W_f 为权重矩阵；h_{t-1} 为上一时刻的隐藏状态；x_t 为当前时刻的输入；b_f 为遗忘门的偏置向量；σ 为 sigmoid 函数。

输入门决定了当前时刻单元状态中保留来自当前时刻输入的程度。它接收上一时刻的 h_{t-1} 和当前时刻的 x_t，通过一个 tanh 函数输出一个取值在(−1, 1)范围内的阈值，从而得到此刻的备选记忆信息 $\tilde{c}_t$。同时，h_{t-1} 和 x_t 通过输入门确定输入的范围，通过遗忘门控制上一时刻的记忆信息的保留。再结合输入门对备选记忆信息的选择，形成当前时刻的新记忆信息 c_t。

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \tag{3}$$

$$\tilde{c}_t = \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right) \tag{4}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t \tag{5}$$

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \tag{6}$$

最后，输出门决定当前时刻的记忆信息有多少输出。h_{t-1} 和 x_t 通过输出门后先确定输出范围，再通过一个 tanh 函数对当前记忆信息 c_t 进行部分选取，经过输出门确定信息输出 h_t。

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \tag{7}$$

$$h_t = o_t \cdot \tanh\left(c_t\right) \tag{8}$$

3. 实证分析

(一) 数据来源

样本量过少会影响模型的预测准确性，为保证有充足数量的样本作为训练集，本文随机选取一只股票——中国银行 2018 年 11 月 16 日到 2023 年 11 月 15 日共计 1213 个交易日的收盘价作为样本数据，收盘价可以作为投资者和交易员制定买卖决策的参考依据。将 1213 个日度数据按照周度进行重采样后，共生成 262 个周度数据值，选取前 252 个数据作为训练集，剩余 10 个数据作为测试集。同时，股票收盘价具有明显的时间序列特征，为观察长期和短期情况下机器学习对股票预测准确程度，选取 ARIMA 模型、SVM 模型和 LSTM 模型进行对比。所有数据均来源于国泰安数据库。

(二) ARIMA 模型训练及结果

(1) 数据预处理。首先对原始数据进行重采样，以周且指定周一为单位求平均值，得到周度数据后，划分测试集与训练集。

(2) 平稳性检验。进行 ADF 检验时，该检验用于评估时间序列数据是否具有单位根，以判断时间序列的平稳性。ADF 检验的假设是，原假设(H0)表明存在单位根，即时间序列是不平稳的；备择假设(H1)表示不存在单位根，即时间序列是平稳的。ADF 结果显示：T 值为−1.680，P 值为 0.441 > 0.05，拒绝原假设，即时间序列不平稳。

(3) 差分处理。对于原始数据为非平稳的时间序列，可通过差分处理使其达到平稳状态。如下图所示，经过一阶差分处理后，序列呈现出更为平稳的趋势，ADF 检验 P 值为 0.00 < 0.05，一阶差分后数据平稳，可以用于 ARIMA 模型。

(4) 模型定阶。确定 p、q 的值有两种方法，一是通过观察 ACF 图和 PACF 图来确定，二是使用 AIC

或 BIC 准则来确定参数。根据 BIC 准则的结果，确定 ARMA 的参数为(1, 1)，由于原始数据是经过一阶差分后变平稳的，所以 d 为 1，即 ARIMA 模型的最优参数为(1, 1, 1)。

(5) 模型训练与预测。输入(1, 1, 1)的参数进行模型训练后，将预测结果与测试集进行对比，MSE 值为 0.0092，MAE 值为 0.0570，R^2 为 0.9195。残差分析、正态分布、QQ 图、相关系数均通过检验。

(三) SVM 模型训练及结果

设置时间窗和特征数据个数，将数据标准化后，同样划分训练集与测试集，从训练集数据中选择最优的超平面，可以是线性或非线性的，以确保划分出的两个类别之间存在最大的间隔距离。经过参数调优后，选择最优参数：10，0.05，rbf进行模型训练与预测拟合，在支持向量机(SVM)模型的参数调优过程中，通过实验和交叉验证，确定最佳参数组合为{'C':10,'gamma':0.05,'kernel':'rbf'}。这组参数的选择是为了在模型性能和泛化能力之间取得平衡，以确保在面对新的未见数据时模型能够表现出色。参数'C'表示软间隔的惩罚项，而'gamma'则影响了核函数的宽度。在最佳组合中，'C'的较高值表明在模型训练中更强调对错误的容忍，从而降低过拟合的风险。同时，较小的'gamma'值有助于确保决策边界的平滑性，防止模型在训练数据中过度拟合。核函数的选择也是调优的关键一步，最终选择的径向基函数(RBF)核，其在非线性数据集上表现出色。这个最佳参数组合代表了在研究的特定问题背景下，SVM模型在训练数据上取得了最佳性能。这样的参数调优过程是为了确保模型在面对实际场景时能够达到最佳的预测效果，兼顾了模型的稳定性和泛化性。SVM模型预测结果的MSE值为0.0040，MAE值为0.0411，R^2为0.6838。

(四) LSTM 模型训练及结果

导入数据后，划分窗口个数和数据特征个数，分别为 7 和 5，输入层的形状是(window_size, fea_num)，其中 window_size 是时间窗口大小，fea_num 是每个时间步的特征数量。模型的结构包括两个 LSTM 层，分别有 128 和 64 个神经元，以及一个 Dense 层，包含 32 个神经元，使用 ReLU 激活函数。输出层具有 1 个神经元，适用于回归问题。在处理数据之前，采用滑动窗口方法，对数据进行归一化，并重新划分为训练集和测试集。选择前 1000 个数据作为训练集，而其余的数据则作为测试集。在模型的编译阶段，使用均方误差(MSE)作为损失函数，并选择 Adam 优化器进行模型训练。同时，衡量模型性能的指标被设置为均方误差。

训练模型后，对数据进行反归一化并预测，LSTM 模型预测结果的 MSE 值为 0.0015，MAE 值为 0.0291，R^2 为 0.9815。

三种模型的预测结果如表 1 所示：

Table 1. Comparison of model prediction results
表 1. 模型预测结果对比

	MSE	MAE	R^2
ARIMA	0.0092	0.0570	0.9195
SVM	0.0040	0.0411	0.6838
LSTM	0.0015	0.0291	0.9815

根据表 1 的结果，可以观察到 LSTM 模型在 MSE 和 MAE 方面均优于 ARIMA 和 SVM 模型。其中，LSTM 模型在 MSE 方面取得最小值，为 0.9815，表明其在拟合准确度上表现出较高水平，预测能力也相对较强。因此，综合考虑各项指标，可以得出 LSTM 模型在预测效果上优于 ARIMA 和 SVM 模型。

4. 研究结论

本文采用 ARIMA 模型、支持向量机(SVM)模型和长短期记忆网络模型(LSTM)，对中国银行 2018 年

11 月 16 日到 2023 年 11 月 15 日共计 1213 个交易日的股票收盘价进行预测，最后通过预测结果对比发现，LSTM 模型的预测结果最好。LSTM 在处理与时间序列密切相关的问题时表现出色，尤其在应对股票数据方面，其模型设计更为契合。相对于其他两种模型，LSTM 能够更有效地预测股票走势。

参考文献

[1] Fama, E.F. (1965) The Behavior of Stock-Market Prices. *The Journal of Business*, **38**, 34-105. https://doi.org/10.1086/294743

[2] Box, G., Jenkins, G.M., Reinsel, G.C., *et al*. (2015) Time Series Analysis: Forecasting and Control. 5th Edition. Wiley, New York.

[3] 闫宇, 吴海涛. 基于 ARIMA 模型的纳斯达克指数短期预测[J]. 信息与电脑, 2020(20): 155-158.

[4] Vapnik, V.N. and Chervonenkis, A.Y. (1968) On the Uniform Convergence of Relative Frequencies of Event to Their Probabilities. *Soviet Math Dokl*, **9**, 915-918.

[5] Muller, K.R., Smola, A.J., Rtsch, G., *et al*. (1997) Predicting Time Series with Support Vector Machines. In: Gerstner, W., Germond, A., Hasler, M., Nicoud, JD., Eds., *Artificial Neural Networks—ICANN*'97. *ICANN* 1997. *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 999-1004. https://doi.org/10.1007/BFb0020283

[6] Pai, P.-F. and Lin, C.-S. (2005) A Hybrid ARIMA and Support Vector Machines Model in Stock Price Forecasting. *Omega*, **33**, 497-505. https://doi.org/10.1016/j.omega.2004.07.024

[7] 阎威武, 邵惠鹤. 支持向量机和最小二乘支持向量机的比较及应用研究[J]. 控制与决策, 2003(3): 358-360.

[8] 顾红其. 支持向量机在期货价格预测中的应用研究[J]. 计算机仿真, 2010, 27(12): 358-360+385.

[9] 张玉, 尹腾飞. 支持向量机在税收预测中的应用研究[J]. 计算机仿真, 2011, 28(9): 357-360.

[10] 周凌寒. 基于 LSTM 和投资者情绪的股票行情预测研究[D]: [硕士学位论文]. 武汉: 华中师范大学, 2018.

[11] 陆泽楠, 商玉林. 基于 LSTM 神经网络模型的钢铁价格预测[J]. 科技视界, 2017(13): 116-117.