E-Commerce Letters 电子商务评论, 2024, 13(2), 2890-2896
Published Online May 2024 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2024.132354

Hans汉斯

基于机器学习的中证500指数期货价格预测

王　奥

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12日；发布日期：2024年5月30日

摘　要

文章采用多种时间序列模型对中证500指数期货收盘价数据展开分析，通过递增窗口交叉验证以及网格调参的方法，系统性地选择最优参数，以提高对期货收盘价数据的预测精度。研究中使用了机器学习中的随机森林、支持向量机、多层神经网络以及ARIMA模型作为基准模型，通过对比分析这些模型的预测效果，从而深入了解它们在对中证500指数期货收盘价时间序列上的性能表现。结果表明：对于该期货价格收盘价的性质和特质，随机森林进行时间序列预测比其他模型的预测精度更高。

关键词

时间序列，机器学习，期货收盘价

Research on the Futures Price of China Securities 500 Index Based on Machine Learning

Ao Wang

School of Economics, Guizhou University, Guiyang Guizhou

Received: Mar. 18th, 2024; accepted: Apr. 12th, 2024; published: May 30th, 2024

Abstract

In this paper, multiple time series models are used to analyze the closing price data of CSI 500 index futures, and the optimal parameters are selected by increasing window cross-validation and network parameters to predict the closing price data. Random forest, support vector machine, multi-layer neural network and ARIMA model in machine learning are used as the benchmark model to compare and analyze the prediction effect. The results show that the time series predic-

文章引用: 王奥. 基于机器学习的中证 500 指数期货价格预测[J]. 电子商务评论, 2024, 13(2): 2890-2896.
DOI: 10.12677/ecl.2024.132354

tion of random forest is more accurate than other models for the nature and characteristics of the closing price of the futures.

Keywords

Time Series, Machine Learning, Futures Closing Price

1. 引言

金融市场的波动性一直是投资者和决策者关注的焦点之一。在现代投资环境中，对于股票和期货价格的准确预测变得至关重要，因为这有助于制定有效的投资和风险管理策略。中证 500 指数作为中国股市的代表性指数，其期货价格的波动对市场参与者具有重要的经济意义。因此，对中证 500 期货价格进行准确的时间序列预测成为金融研究中的一项重要任务。本研究的动机在于通过比较不同类型的预测模型，包括传统的 ARIMA 模型、支持向量机(SVM)、随机森林和多层神经网络(MLP)等，来寻找一种更为准确和可靠的中证 500 期货价格预测方法。本研究的主要目的是探究不同模型在中证 500 期货价格预测中的表现差异，并为金融决策者提供更可靠的市场趋势预测信息。通过比较模型在 MSE (均方误差)等性能指标上的表现，本文旨在确定哪种模型在这一特定场景中表现最佳，以指导实际的金融决策。本研究采用了多种时间序列预测模型，包括传统的 ARIMA 模型、机器学习中的支持向量机(SVM)、随机森林和深度学习中的多层神经网络(MLP)。通过使用历史中证 500 期货价格数据，本文训练并评估了这些模型的性能。评估指标主要包括 MSE，以量化不同模型的预测精度。

本研究的贡献在于提供了对于中证 500 期货价格预测的全面分析，深入比较了不同类型模型并加上其组合模型的性能。通过揭示每种模型的优势和局限性，期望为金融市场参与者提供更明晰的决策支持，促进更精准的投资和风险管理策略的制定。研究的结果对于金融领域的学术研究和实际应用都具有一定的指导意义。

2. 文献综述与模型介绍

2.1. 文献综述

期货价格预测是金融、经济等领域的研究热点之一，而期货数据是一类具有代表性的金融数据，近年来吸引了众多研究人员的关注。闫宇、吴海涛(2020) [1]利用纳斯达克综合指数(NASDAQ)的数据，通过平滑化的 ARIMA 模型对股市的短期趋势进行了预测。吴玉霞(2016) [2]利用 ARIMA 模型对我国上市公司的股价走势进行了分析。通过对模型的短期动静态预测，验证了模型的有效性。隋学深(2008) [3]利用上海股市各个尺度系数的趋势性与记忆性，利用 SVM 方法对股市进行了预测，结果表明该模型对股市走势的预测准确率大于 50%，而 SVM 对股市的分类准确率达到了 60%，证实了 SVM 应用于股市预测的有效性。李志杰(2015) [4]利用神经网络和主成分分析相结合的方式，对原始数据进行降维，实验证明，通过对原始数据的降维，得到了较好的学习效果。基于 SVM、贝叶斯以及随机森林模型，周翔等(2020) [5]分别对信贷违约的进行了预测，并以准确率、AUC 和漏警率三个指标来比较分析不同模型的预测效果，实证结果显示随机森林模型具有最为优异的预测性能表现。GHOSH *et al.* (2022) [6]采用 RF 和 LSTM 作

为训练算法，证明了它们在预测标普成分股价格定向变动方面的有效性。

基于现有研究，本文通过递增窗口交叉验证和网格调参在几种时间序列预测模型中选取最优人工智能模型预测中证 500 指数的期货收盘价。首先，组合多种时间序列模型，将随机森林、支持向量机、多层神经网络和 ARIMA 模型组合在一起，形成一个综合的预测框架。这种组合可以弥补单一模型的局限性，提高整体预测效果。其次，进行递增窗口交叉验证，采用递增窗口交叉验证方法，能够更有效地评估模型的泛化能力，并且在训练过程中动态调整模型参数，使其更好地适应不同时间段的数据特征变化。并对它们的预测效果进行对比分析。通过深入比较这些模型在预测期货收盘价方面的表现，可以揭示各自的优势和劣势，为未来的研究提供参考。综上所述，文章的创新性工作在于将多种时间序列模型进行组合，并采用递增窗口交叉验证和网格调参优化方法，从而提高对期货收盘价数据的预测精度，并对不同模型的效果进行了全面比较分析。

2.2. 模型介绍

1) 随机森林模型

随机森林是一种基于集成学习的分类和回归算法。随机森林通过构建多个决策树，并对它们的结果进行综合，以提高整体模型的准确性和稳定性。随机森林首先从原始训练集中随机选择部分样本进行有放回抽样。这样会形成多个大小相似的训练集，用于构建每个决策树。接着对于每个决策树的每个节点，随机选择一部分特征子集。这样可以减少特征之间的相关性，并增加决策树之间的多样性。然后基于随机采样和随机特征选择的训练集，构建多个决策树。每个决策树都通过对特征空间进行划分来进行训练，直到达到终止条件，例如达到最大深度或节点中的样本达到最小数量。

2) SVM 模型

支持向量机(Support Vector Machine，简称 SVM)是一种监督学习算法，常用于分类和回归分析。SVM 的核心理论基础是在特征空间中寻找一个最优超平面，能够有效地将不同类别的样本分开。

SVM 的基本思想是将样本表示为特征空间中的点，其中每个样本被赋予一个标签。目标是找到一个超平面，该超平面能够将不同类别的样本分开，并且与最近的样本之间的间隔最大化。这些最近的样本点被称为支持向量，因为它们对于定义分类决策边界非常重要。

SVM 的核心概念在于核函数，这一关键概念使其能够在低维特征空间中进行非线性分类。核函数的主要作用是通过将数据映射到高维空间，实现在高维空间中线性可分的效果。SVM 中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) 多层神经网络模型

多重神经网络系统(Multi-Layer Neural Network Systems)是一种深度学习模型，也称为深度神经网络(Deep Neural Networks, DNN)。它由多个神经网络层组成，每一层都由多个神经元组成。这些层之间的连接是前向传播的，每个神经元通过激活函数将输入信号传递到下一层。多层神经网络系统的核心思想是通过逐层处理数据来学习复杂的特征表示。每一层都对输入数据进行非线性变换和特征提取，以便更好地捕捉输入数据中的抽象特征。底层的网络层负责捕捉低级特征，而较高层的网络层则通过组合底层特征来构建更高级别的特征表示。深度神经网络系统的训练过程通常使用反向传播算法来优化网络权重，以最小化的预测输出与实际标签之间的误差。通过增加网络的深度和参数数量，深度神经网络能够更好地捕捉数据中的复杂结构和模式，提高模型的性能和准确性。

4) ARIMA 模型

ARIMA 是一种经典的时间序列预测方法，结合了自回归(AR)和移动平均(MA)的概念。ARIMA 模型通常用于对非平稳时间序列数据进行建模和预测。ARIMA 模型由三个部分组成：自回归(AR)部分、积分

(I)部分和移动平均(MA)部分。具体来说，ARIMA (*p*, *d*, *q*)模型的定义如下：

AR (自回归)部分：该部分表示当前观测值与过去观测值之间的关系，其中 *p* 是自回归项的阶数。

$$\left[Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t\right] \tag{1}$$

其中，(Y_t)是时间点(t)处的观测值，$\left(\phi_1, \phi_2, \cdots, \phi_p\right)$是自回归系数，$\left(\epsilon_t\right)$是白噪声误差。

I (积分)部分：该部分表示为使时间序列变得平稳而进行的差分操作，其中 *d* 是差分的次数。

$$\left[\text{diff}\left(Y_t, d\right) = \left(1 - B\right)^d Y_t\right] \tag{2}$$

其中，(B)是滞后算子，(diff (Yt, d))表示对(Y_t)进行(d)阶差分。

MA (移动平均)部分：该部分表示当前观测值与过去白噪声误差的线性组合，其中 *q* 是移动平均项的阶数。

$$\left[Y_t = \epsilon_t + \theta_1 \epsilon_{(t-1)} + \theta_2 \epsilon_{(t-2)} + \cdots + \theta_q \epsilon_{(t-q)}\right] \tag{3}$$

其中，$\left(\theta_1, \theta_2, \cdots, \theta_q\right)$是移动平均系数。

2.3. 模型调优

在机器学习中，通常都是通过 k 折交叉检验来选取最佳的模型和参数[7]，但是，如果将这种方法用于时间序列，那么就会产生“用未来的信息来预测现在的信息”的局面，而通过按时间序列的增窗进行交叉验证，则能够确保没有用来对当前的信息进行预测。

本文将原始样本分为训练样本和测试样本，然后对样本进行训练，然后用增窗交叉验证和网络方法选择最优模型，然后对样本进行一次外推，并对各模型的预测结果进行比较。将训练数据划分为 N + 1 个子集，每个子集的样本数为训练集总样本数除以(N + 1)。将所得余数放入第 1 个子集作为第一期训练样本(Train1)，而将第 2 个子集到第 N + 1 个子集作为验证集(Valid1)。按照这一规则划分，每个验证集的样本数相同。

3. 实证分析

3.1. 数据来源

本文选取了 2018 年 1 月 1 日至 2023 年 12 月 1 日的中证 500 指数为标的的 20_IF0Y00 期货合约收盘价日度数据，在 RESET 数据的金融期货数据库获取。中证 500 指数期货是以中证 500 指数为标的物的金融衍生品。期货合约是一种衍生工具，其价格取决于标的物(在这里是中证 500 指数)未来的表现。期货市场为投资者提供了对冲风险、进行套期保值、实现杠杆交易等机会。对数据进行异常值和缺失值处理，描述性统计如下表 1 所示。偏度衡量了数据分布的不对称性。正偏度表示分布右偏，负偏度表示分布左偏。在这里，偏度接近于零，表示数据分布相对对称。峰度衡量了数据分布的尖峭度。负峰度表示比正态分布稍微平缓。

Table 1. Descriptive statistics
表 1. 描述性统计

个数	均值	标准差	最小值	最大值	偏度	峰度	标准误	JB 检验值
1437	4183.40	586.01	2962.2	5901.0	0.4746	−0.5124	15.46	69.72***

***为 1%显著水平。

3.2. 模型调优

首先，本文对该日度数据进行归一化处理(Min-Max scaling)，将所有数据缩放到[−1, 1]区间，其中x_i代表第 i 个特征数据，x_{min}为特征数据的最小值，x_{max}为最大值。归一化公式如下所示：

$$x = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{4}$$

最后选取 30 个交易日数据作为测试集，其他数据作为训练集。

1) RF 模型最优参数选择

最大深度：设置为 5，表示每棵决策树的最大深度为 5。这是一种对树进行剪枝的方法，有助于控制模型的复杂度和防止过拟合。最大特征数设置为“auto”，表示在寻找最佳分割时考虑所有特征。这是典型的随机森林做法，它引入了随机性以提高模型的泛化性能。叶节点最小样本数设置为 2，表示每个叶节点最少包含 2 个样本。这可以控制叶子节点的大小，有助于防止模型在训练集上过分拟合。内部节点最小样本数设置为 2，表示在内部节点分裂时所需的最小样本数。这个参数同样有助于控制树的生长，防止过拟合。树的数量设置为 200，表示随机森林中包含 200 棵决策树。增加树的数量可以提高模型的鲁棒性和泛化性能，但也增加了计算成本。

2) SVM 模型最优参数选择

惩罚项参数 C 设置为 1，这是惩罚项的强度。较小的 C 值表示较强的正则化，对误分类的惩罚力度较小。kernel 设置为“linear”，表示使用线性核函数。线性核适用于线性可分的数据集，它在特征空间中进行线性分割。

核函数参数 gamma 设置为 0.01，这是核函数的系数。对于线性核函数，gamma 的值通常不太重要，因为它在该情况下不会被使用。在使用非线性核函数时，gamma 控制了样本点影响的范围，较小的 gamma 值表示影响范围较大。

这个参数选择表明了一个相对简单的 SVM 模型，它在特征空间中使用线性分割，并具有一定的正则化(由 C 控制)。这样的模型可能更适用于线性可分的问题，而且在处理较大的数据集时，较小的 C 值可以防止过拟合。

3) 多层神经网络模型最优参数选择

激活函数 activation 设置为“logistic”，表示使用逻辑斯蒂回归函数作为激活函数。逻辑斯蒂函数可以将输出限制在[0, 1]范围内，常用于二分类问题。正则化参数 alpha：设置为 0.0001，这是正则化项的权重系数。较小的 alpha 值表示较弱的正则化，可以减少过拟合的风险。隐藏层大小 hidden_layer_sizes 设置为(100, 50)，表示共有两个隐藏层，第一个隐藏层有 100 个神经元，第二个隐藏层有 50 个神经元。隐藏层的大小决定了神经网络的复杂度和表示能力。迭代次数 max_iter：设置为 5000，表示训练时的最大迭代次数。这个参数控制了神经网络的训练过程的总体迭代次数。优化器 solver：设置为“adam”，表示使用 Adam 优化器进行参数优化。Adam 优化器结合了 AdaGrad 和 RMSProp 的优点，并能有效地处理大型数据集和高维空间。

4) ARIMA 模型最优参数选择

ARIMA 模型在时间序列数据分析及预测方面具有广泛应用。见表 2，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。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 (见表 3)，处理后数据平稳，可以用于 ARIMA 模型。ARIMA (p, d, q)模型中，p 表示自回归的滞后阶数，d 表示序列单整阶数，q 表示移动平均滞后阶数。在可视化 ACF 和 PACF 之后，本文还通过 AIC 和 BIC 最小值来确定 p，q 值，最后确定 p，d，q，然后使用 ARIMA (1, 1, 0)来拟合收盘价数据。

Table 2. ADF test of the original sequence
表 2. 原序列 ADF 检验

	T-statistics	P 值
收盘价	ADF Statistic: −1.5668	0.5002
	1%: −3.4349	
	5%: −2.8635	
	10%: −2.5678	

Table 3. ADF test after first-order difference
表 3. 一阶差分后 ADF 检验

	T-statistics	P 值
收盘价	ADF Statistic: −38.3767	0.00
	1%: −3.4349	
	5%: −2.8635	
	10%: −2.5678	

3.3. 模型性能判别

均方根误差(Root Mean Square Error，简称 RMSE)是一种衡量预测模型误差的指标。它通过计算预测值与实际观测值之间的差异，然后取平方、求平均并开方来得到一个标准化的误差值。RMSE 通常用于评估回归模型的性能，越小的 RMSE 值表示模型对观测值的拟合越好。公式如下：

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}} \tag{5}$$

本文的结果如表 4 所示，所有的时间序列预测模型皆通过 PT 测试，且在 1%的置信水平上显著，模型的预测值与真实值间的相对变化都是同步的，但预测数据仍呈现一定的延后性。在所有的模型中随机森林的均方根误差最小，多层神经网络最大，RF 在对中证 500 指数期货收盘价的预测效果相对于其他几个模型是最佳的。

Table 4. Comparison of model predictions
表 4. 模型预测比较

模型	RMSE 值	PT 测试
ARIMA	0.000301	***
RF	0.000182	***
SVM	0.000250	***
多层神经网络	0.000326	***
ARIMA + RF	0.000242	***
ARIMA + SVM	0.000276	***
ARIMA + 神经网络	0.000319	***

公式 1 预测值的相对变化与真实值的相对变化是同步的。在 1.00%的显著性水平上未能拒绝原假设。

4. 研究结论

通过对证 500 期货合约收盘价时间序列的预测性能进行比较，在本研究中，随机森林模型表现出色，其拟合效果明显优于其他模型。随机森林的强大非线性建模能力和对复杂数据关系的适应性使其在时间序列预测任务中表现突出。其次，支持向量机模型展现了相对较好的性能。SVM 以其在高维空间中的强大分类和回归能力而闻名，对于期货合约收盘价的预测显示出良好的表现。其在处理非线性关系方面的优势使其成为时间序列分析中的一种有力工具。ARIMA 模型在时间序列分析中有着广泛的应用，然而，相较于机器学习方法，其对于复杂数据模式的捕捉能力相对较弱，因此在本研究中表现居中。多层神经网络，尽管被广泛应用于各种预测任务，但在本研究中表现相对较差。这可能是由于神经网络对于数据量的敏感性以及需要仔细调整的超参数而导致的。对于期货合约收盘价时间序列的特定问题，其他模型似乎更为适用。总体而言，本研究的结果强调了在金融时间序列预测中，机器学习方法，特别是随机森林和支持向量机，可能更具优势。然而，最佳模型的选择仍然取决于具体问题的性质和数据的特征，未来的研究可以进一步探索不同模型在更广泛金融市场情境下的应用。

参考文献

[1] 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.

[2] 吴玉霞, 温欣. 基于 ARIMA 模型的短期股票价格预测[J]. 统计与决策, 2016(23): 83-86.

[3] 隋学深, 齐中英. 基于多尺度特征和支持向量机的股市趋势预测[J]. 哈尔滨工业大学学报(社会科学版), 2008(4): 77-82.

[4] 李志杰. 基于神经网络的上证指数预测研究[D]: [硕士学位论文]. 广州: 华南理工大学, 2015.

[5] 周翔, 张文宇, 江业峰. 个人信贷违约预测模型的研究[J]. 辽宁科技大学学报, 2020, 43(3): 223-230.

[6] Ghosh, P., Neufeld, A. and Sahoo, J.K. (2022) Forecasting Directional Movements of Stock Prices for Intraday Trading Using LSTM and Random Forests. *Finance Research Letters*, **46**, Article ID: 102280. https://doi.org/10.1016/j.frl.2021.102280

[7] Ince, H. and Trafalis, T.B. (2008) Short Term Forecasting with Support Vector Machines and Application to Stock Price Prediction. *International Journal of General Systems*, **37**, 677-687. https://doi.org/10.1080/03081070601068595