摘　要

股票市场是社会主义市场经济体系的重要组成，股票价格的波动可以反映宏观经济的发展状况。随着市场交易的日常化，股票价格预测逐渐成为当前学界业界共同关注的问题之一。然而，股票价格受着多因素的影响，时刻处于非线性的动态变化之中。若能实现较为精确的预测，将有助于投资者降低投资风险、建立组合投资策略，也将为中国股市的理论研究提供有力参考。因此，如何选择和量化股票价格预测的影响因素、应用何种预测模型，便具有重要的理论研究价值。

鉴于股票价格关于其影响因素变化的敏感性，其预测难度巨大。人们不断尝试应用包括历史交易信息、宏观经济指标、技术指标、互联网舆情和金融研报等各式各样的数据源，以及应用诸如现代计量经济学、机器学习和人工智能等各类新的模型方法，进行股票价格预测研究。不同来源的数据影响角度不同，融合多种数据源的股票价格预测研究，可充分发挥数据间的关联信息，提升预测精度。为此，本文以中国平安（601318.SH）为研究对象，通过变量设计融合其历史交易信息、基本面特征、技术特征以及情感特征这四种数据源，在原有 43 个预测变量的基础上生成了 15 个特征向量，以此为基础开展股票价格的预测研究。

为了规避高维数据带来的维度灾难风险，本文首先运用主成分分析法（Principal Components Analysis，PCA）对输入数据进行降维，以消除数据冗余，提高模型深度学习效率；其次引入自适应噪声完备集合经验模态分解方法（Complete Ensemble Empirical Mode Decomposition with Adaptive Noise，CEEMDAN）对前收盘价价格序列进行分解，并利用由粗到细的均值重构算法生成不同尺度的新特征，以充分挖掘价格序列中的隐含信息，缓解预测常存的滞后问题。然后，将主成分得分与重构后的新特征融合，初步构建了股票价格预测长短期记忆神经网络模型（Long Short-Term Memory，LSTM），理论上为模型泛化能力增强和预测效果提升奠定基础。

LSTM 深度学习算法对于长期时间序列的处理能力较强，但在应用 LSTM 训练学习之前，需要对诸如隐藏层神经元数量、dropout 比率、时间步长等相关控制参数进行赋值。为了实现 LSTM 网络模型的最佳性能，本文引入粒子群优化算法（Particle Swarm Optimization，PSO），实现对 LSTM 控制参数的智能寻优，

以确保输入特征与网络单元结构相匹配，进一步完善了融合多源数据的PCP-LSTM股票价格预测模型。

最后，本文设置了两组对照实验以验证所构建的预测模型是否具有有效性：一是模型对照组，即与其他深度学习模型的预测效果进行对比；二是数据源对照组，即比较PCP-LSTM模型在不同数据集上的预测表现。此外，本文还通过两种方式检验了PCP-LSTM模型的稳健性：一是通过LSTM算法滑动扩充出测试集以外的新数据集并利用PCP-LSTM模型进行样本外预测；二是基于相同方法的变量设计并利用PCP-LSTM模型对具有代表性的上证综合指数进行预测。实验结果表明，本文所构建的PCP-LSTM股票价格预测模型具有最佳预测性能和一定的泛化能力和稳健性，并且基于多源数据融合具有实际意义。

关键词：股票价格预测；PCP-LSTM模型；样本外预测；多源数据融合

Abstract

The stock market is an important component of the socialist market economy and the fluctuation of stock prices can reflect the development of the macro economy. With the daily trading of the market, stock price forecasting has gradually become one of the common concerns of the current academic community. However, stock prices are influenced by many factors and are constantly in a non-linear and dynamic state. A more accurate forecast would help investors to reduce investment risks and build portfolio investment strategies, and would also provide a strong reference for theoretical research on the Chinese stock market. Therefore, the selection and quantification of the factors influencing stock price forecasting and the application of forecasting models are of great theoretical value.

Given the sensitivity of stock prices to changes in their influencing factors, their prediction is extremely difficult. Attempts have been made to apply a variety of data sources, including historical trading information, macroeconomic indicators, technical indicators, internet opinion and financial research reports, as well as new modelling methods such as modern econometrics, machine learning and artificial intelligence, to conduct research on stock price forecasting. Different sources of data have different perspectives of influence, and stock price forecasting studies that integrate multiple sources of data can make full use of the correlation information between data and improve forecasting accuracy. To this end, this paper takes Ping An of China (601318.SH) as the research object and uses its four data sources - historical trading information, fundamental characteristics, technical characteristics and sentiment characteristics - to generate 15 feature vectors based on the original 43 predictor variables as a basis for stock price prediction research.

In order to avoid the risk of dimensional disaster caused by high-dimensional data, this paper firstly uses Principal Components Analysis (PCA) to dimensionally reduce the input data to eliminate data redundancy and improve the model's deep learning efficiency; secondly, it introduces the CEEMDAN signal decomposition method to decompose the previous closing price series, and uses fine to coarse

reconstruction algorithm to generate new features at different scales to fully exploit the implicit information in the price series and alleviate the lag problem that often exists in forecasting. Then, the principal component scores are fused with the reconstructed new features to construct a Long Short-Term Memory (LSTM) neural network model for stock price prediction, which theoretically lays the foundation for enhancing the generalization ability and prediction effectiveness of the model.

The LSTM deep learning algorithm is more capable of handling long-term time series, but before applying the LSTM training learning, relevant control parameters such as the number of hidden layer neurons, dropout ratio, time step and so on need to be assigned. In order to achieve the best performance of the LSTM network model, this paper introduces the Particle Swarm Optimization (PSO) algorithm to achieve intelligent optimisation of the LSTM control parameters to ensure that the input features match the network cell structure, further improving the PCP-LSTM stock price prediction model incorporating multiple sources of data.

Finally, two sets of control experiments are set up to verify the validity of the constructed prediction models: a model control group, which compares the prediction performance with other deep learning models; and a data source control group, which compares the prediction performance of the PCP-LSTM model on different datasets. In addition, the robustness of the PCP-LSTM model is tested in two ways: first, by sliding the LSTM algorithm to expand a new dataset outside the test set and using the PCP-LSTM model to make out-of-sample predictions; second, by using the PCP-LSTM model to predict a representative SSE Composite Index based on the variable design of the same method. The experimental results show that the PCP-LSTM stock price forecasting model constructed in this paper has the best forecasting performance and certain generalization ability and robustness, and is practically meaningful based on multi-source data fusion.

Key words： Stock Price Forecasting; PCP-LSTM Model; Out-of-sample Predictions; Multi-source Data Fusion

目 录

第 1 章　引言

1.1　选题背景和意义

1.1.1 选题背景

伴随着市场经济与社会经济的快速发展，股票愈发受到投资者的关注，中国股民数量逐年增加，根据中国证券登记结算有限责任公司发布的数据，截至 2022 年 3 月，全国股票投资者达 20244.94 万，其中自然人投资者 20196.91 万，占比高达 99.76%。图 1-1 显示了 2020-2022 年上证综合指数的日收盘价价格和涨跌幅情况，可以看出数据总体呈现出非线性趋势且波动程度较为明显，无特殊规律可寻。因此，在巨大的投资金额和复杂的股票价格变动趋势的双重压力下，建立相对准确的模型实现价格预测，对于专业知识匮乏、风险承受能力较差、普遍具有从众心理的大多数投资者以及中国股市的健康发展均具有重要作用。

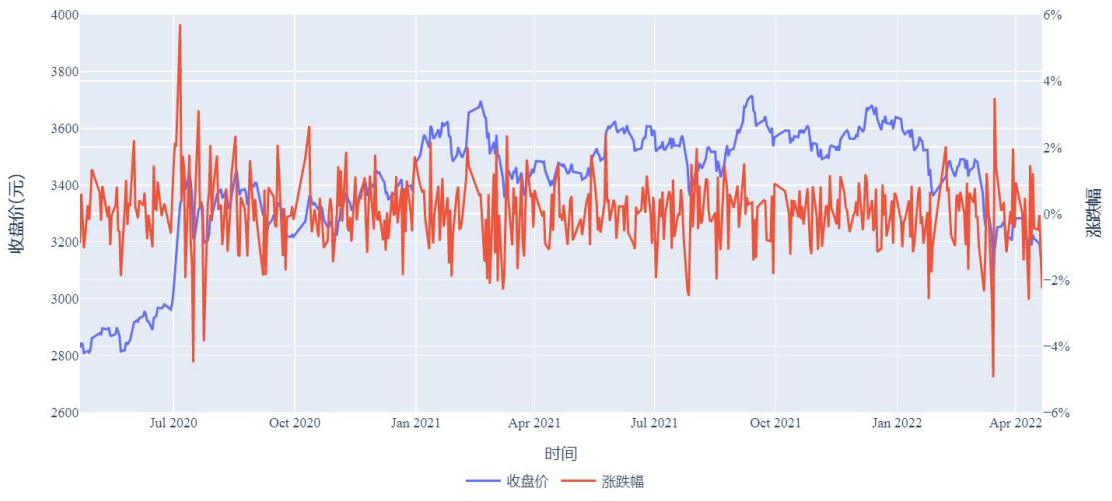

图 1-1 **2020-2022 年上证综合指数日收盘价与涨跌幅变动趋势**

对于股票价格这种时间序列数据，研究者较早主要采用计量模型和数学模型进行预测，常见的有自回归移动平均模型（AutoRegressive Moving Average Model，ARMA）、差分整合移动平均自回归模型（AutoRegressive Integrated Moving Average model，ARIMA）、广义自回归条件异方差模型（Generalized AutoRegressive Conditional Heteroskedasticity，GARCH）或是马尔可夫模型（Markov Model，MM）等。然而，计量模型常需要序列是线性平稳的，数学模型在问题求解时常需要进行纷繁的计算甚至是计算机的模拟，因此在面对影响因素多、非线性的股票数据时，两类模型的局限性较高，应用性不强。随着计算机科学技术的发展，越来越多的研究者在预测股票价

格时逐渐进入到机器学习这一行列中，如灰色预测模型（Grey Models，GM）、支持向量机（Support Vector Machine，SVM）、随机森林（Random Forest，RF）、BP神经网络等，虽然它们都可以在一定程度上处理非线性的数据，但对高维数据、时间序列数据处理能力有限、预测效果欠佳。近些年，计算机处理能力迅速提升，深度学习在机器学习的基础上更进了一步，能够以更高的精度、更少的时间以及更简便的操作得到更好的预测效果，故而成为了股票价格预测的新方向。其中，卷积神经网络（Convolutional Neural Network，CNN）与循环神经网络（Recurrent Neural Network，RNN）作为深度学习的两大代表算法较早被提出，实现了股票数据时间序列特征的深层次捕捉。但CNN没有记忆功能且在应用过程中池化层会丢失部分价值信息，从而对预测性能造成了直接影响；RNN在求解损失函数的最小值时容易出现梯度消失问题，导致了对于时间跨度较大的序列缺乏记忆性；而长短期记忆神经网络（Long Short-Term Memory，LSTM）和门控循环单元神经网络（Gate Recurrent Unit，GRU）通过引入“门”的概念，实现有选择的记忆与遗忘信息，从而很好地解决了记忆性问题。

在对股票价格预测时，容易遇到以下两个问题：其一，非线性系统内部运行机制较为复杂，难以深入理解，建模难度较大；其二，股票价格影响因素较多、特殊性较强，在选择输入变量时不易识别哪些是重要的解释变量，哪些是不显著变量或无关变量，加之诸如政治、经济、心理等影响因素我们无法直接量化，从而导致模型精度较低，实现较为准确的预测相对困难。

针对上述情况，本文将尽可能多地融合来自不同数据源的数据以弥补影响因素选取不充分进而预测困难的问题，并通过降维方法在保留大部分信息的同时抓住主要矛盾以揭示事物内部变量之间的规律性；采用粒子群优化算法（Particle Swarm Optimization，PSO）帮助深度学习模型发挥更好的性能，以处理棘手的非线性问题、弥补传统机器学习方法中的不足、挖掘更深层次的信息，从而以更好的模型获得更准确的预测。

1.1.2 研究意义

（1）理论意义。本文所构建的股票价格预测模型是在依据对输入指标及预测模型的已有研究下，在深度学习的基础上融合相关方法而成，以检验模型对于非线性数据的处理与预测能力，使其在理论上具备可行性。利用主成分分析法（Principal

Components Analysis，PCA）对输入变量降维，以减少信息冗余，增加学习速率；采用自适应噪声完备集合经验模态分解（Complete Ensemble Empirical Mode Decomposition with Adaptive Noise，CEEMDAN），防止损失蕴含在不同频率下的价格变化信息；应用 PSO 算法为模型寻求最优参数，增加预测性能。上述方法与深度学习算法的有机结合为股票价格的研究开创了新的尝试思路。

（2）现实意义。研究的实际应用价值主要体现在通过把握中国市场的股价变动情况，从而为投资者的决策提供参考。在当前的市场环境下，信息不对称是投资者面临的最大问题，因此，对于广大投资者来说，掌握股票价格预测背后的思维方法和逻辑路径至关重要，例如，如何分析股票价格走势？分析时使用了哪些维度，应用了何种方法？通过对股票价格预测的研究，不仅可以有效弥补信息获取不及时、不充分的问题，而且还有助于投资者与时俱进自己的思维方式，从而形成自己独有的投资标准，最终在未来的股市中获得更高的胜率。

1.2　文献综述

1.2.1 股票市场可预测性研究

随着经济体制的不断改革和深入，市场经济日渐活跃，大量投资者涌入股票交易市场，久而久之，股票成为了投资者日常生活中必不可少的一部分。为在追求股票高收益性的同时不冒较高风险，其价格预测便成为了学术界和投资者关注的热点问题。

对于金融资产价格是否可以预测，学术界意见不一。Fama（1965，1970）所提出的有效市场假说（Efficient Markets Hypothesis，EMH）是早先最具有代表性的观点，其中指出，在一个强式有效市场中，无论何时，股价都是有关该股票所有已知信息的充分体现[1-2]。由于所有的已知信息均可以被市场参与者最优地使用，新信息的发生是随机的，因此，股票价格的变化表现为“随机漫步”过程，投资者不可能跑赢市场[3]。然而，与此相对应的行为金融学理论认为，投资者的心理、情绪在很大程度上会影响证券市场的价格变动，加之市场信息的不对称性，从而为股票市场的规律形成提供了空间，存在了价格预测的可能（Kahneman etc，1979；Bondt etc，1985；Jegadeesh etc，1993）[4-6]。基于学术界当中有效市场假说与行为金融学之间的长期争论，创新型观点层出不穷[7]。Lo 将有效市场假说与行为金融学的合理观点进行了有机结合，认为市场有效性与可预测性并不是互相冲突的，进而提出适应性市场假说（Adaptive

Market Hypothesis，AMH），强调股票价格具有可预测性，并且可以从中获利（Lo，2004；Lo，2017）[8,9]。

大量研究表明，国内股票交易市场不满足弱式有效假说（Cao etc，2011；崔宾阁等，2012）[10,11]，市场对信息披露的监管力度不足，使得各投资者之间、投资者与上市公司之间存在信息不对称的问题，且这一过程具有持续性，股票价格因此会受到波动，存在了可预测的成分[12]，从而为接下来进一步的建模研究提供了可能。但是，股票价格往往会受到政治、经济、社会、心理等因素的复杂影响[13-15]，对以上信息变化特别敏感，而这些信息往往很难用变量衡量甚至无法获取。因此，建立股票价格预测模型，保证模型的精度与泛化能力是一项值得探索的工作。

1.2.2 股票价格预测变量研究

建立一个好的股票价格预测模型，输入特征的选择是首要任务，更是成功的前提。因此，选取何种类型的输入特征以及如何选取值得我们的深入研究。综合前人的文献，可将股票价格预测模型的输入特征大致分为四种类型（Chen etc，2015；饶东宁等，2017）[12,16]：

（1）历史价格数据：如开盘价、最高价、最低价、成交量、成交金额等。

（2）技术特征：由历史价格数据计算而得，常见的有简单移动平均线（*SMA*）、指数移动平均线（*EMA*）、指数平滑异同平均线（*MACD*）、相对强度指标（*RSI*）、随机指标（*KDJ*）、威廉指标（*W%R*）和变化率（*ROC*）等[12,17]。

（3）基本面特征：是指反映宏观经济运行情况与行业发展质量的指标，以衡量股票的内在价值。例如市盈率、市净率、市现率、市销率、净资产收益率等。

（4）情感特征：基于情感词典[18,19]或机器学习方法[20-22]对获取的股票评论数据进行文本分析，以得出量化投资者心理情绪的相关指标。

通过对比研究发现，在股票价格预测当中，无论使用传统的计量经济模型还是机器学习模型，大多数文献都没有考虑到股票价格是众多影响因素综合作用的结果，只是片面地利用股票公开的历史价格数据进行建模（郑睿颖和伍应环，2011；陈嶷瑛等，2014；陈卫华，2018；宋刚等，2019；包振山等，2020；Budiharto etc，2021；Kavinnilaa etc，2021）[3,23-28]，预测结果不仅存在严重的滞后性，而且偶然性强、精度低，模型的推广能力也得不到保证。伴随着研究模型的不断改进及完善，研究者们开始寻求新的突破口，尝试引入其他数据源的变量作为预测模型的输入特征，其中，常见的研究

主要包括技术特征（綦方中等，2020；王东等，2021；胡聿文，2021）[29-31]、基本面特征（裴大卫和朱明，2019；闫政旭等，2021；曹超凡等，2021）[32-34]、情感特征（陈晓红等，2016；毛月月和张秋悦，2020；Shahi etc，2020）[20,21,35]中的单一数据源引入,以及两种数据源的混合引入（Beyaz etc，2018；康瑞雪等，2021；赵丽君等，2021；李潇俊和唐攀，2022）[36-39]。已有文献中，多数选取了股票的历史价格数据和技术特征作为模型的输入变量进行预测。除此之外，还有极少数文献选取的相关指标可以覆盖到上述数据源中的三种[40]。大量研究表明，增加输入特征可以使模型在充分的股票市场信息中更好地抓住股票价格变化的内在特点，以实现模型对股票价格预测能力和泛化能力的大幅提升（Chen etc，2015；Roondiwala etc，2017；马超群等，2021）[16,41,42]。

1.2.3 股票价格预测方法研究

总结现有文献，股票价格预测的研究方法主要可以分为以下三种类型：

（1）通过建立计量模型预测的方法

以计量经济学为理论基础建模预测，多为线性模型，例如 ARMA 模型、ARIMA 模型、GARCH 模型以及各种模型的混合模型（程昌品等，2012；杨琦和曹显兵，2016；Rounaghi etc，2016；Herwartz，2017）[43-46]。而实际生活中，股票价格数据往往是非线性、非平稳的，利用线性模型建模不能充分挖掘出股票价格的内在变化特点，模型结果虽然可以得到较好的解释但研究假设具有局限性，预测结果通常不具有信服力[47]。

（2）传统的机器学习方法

传统机器学习方法的应用成功为股票价格预测开创了一片新天地。例如，贝叶斯学习（Du etc，2016）[48]、决策树（李想，2017；衣静，2020）[49,50]、支持向量机（MacKinnon etc，2015；张佩奇，2020）[51,52]、随机森林（闫政旭等，2021）[33]，这些方法虽然一定程度上克服了线性模型的缺陷，但是在学习过程中对样本分布的要求较为严格，导致模型预测精度较低，且泛化能力不足[53]。除此之外，BP 神经网络虽然也属于传统的机器学习方法，但适用领域更广、模型研究更深。蔡红和陈荣耀分别建立了 BP 神经网络、ARMA 模型和线性模型对上海证券交易所上市的首创股份的最高价进行预测，结果表明，BP 神经网络的预测能力最佳[13]；崔文喆等将上海 A 股 30 支股票的收盘价作为模型的预测变量，发现随着预测周期的增长，BP 神经网络的预测精度相较于 GARCH 模型越来越高[54]。BP 神经网络的优势在于，训练过程中，仅需对隐

藏层的数量、阈值和权值进行合理调节，就可以使模型轻松达到以任意精度逼近任何一个连续函数的能力，从而有效地降低预测误差，提升模型的使用价值。但与此同时，BP 神经网络也存在收敛速度慢、容易陷入局部极小值等缺点，股票价格的预测精度也因此受到了一定限制[55]。相关文献表明，传统的机器学习方法对于非线性的时间序列数据建模能力有限，并且当输入特征的维度较高时，学习效果较差，容易引发维数灾难致使模型预测结果出现异常，严重影响了模型的拟合效果。

（3）深度学习方法

随着科学技术的突飞猛进，股票价格的预测方法呈现出向深度学习转变的趋势。常见的深度学习方法主要有 RNN（乔若羽，2019）[56]、CNN（Tsantekidis，2017；Chen etc，2019）[57,58]、LSTM、GRU 等。其中，LSTM 与 GRU 作为 RNN 的变种形式，解决了 RNN 在训练过程中的梯度消失和梯度爆炸问题。大量研究结果表明，深度学习与传统的机器学习方法相比具有良好的自学习、自组织、自适应能力[59]，可以更好地提取和分析隐藏在股票价格数据中的关联特征,模型预测性能较强。然而，在深度学习常用的方法中，不同模型对于时间序列数据的适应性也会存在差异[60]，例如，RNN、CNN 等模型在实际应用过程中对于时间序列数据不够敏感，往往容易忽略其趋势特征；相比之下，LSTM 因其特殊的门控机制呈现出良好的选择性以及有效的长期记忆性，更适合处理这类时间序列数据[32,61]。

虽然 Hochreiter 和 Schmidhuber（1997）较早就已指出 LSTM 模型在语言识别、人机交互等领域，预测能力远胜于 BP 神经网络等模型，但 LSTM 真正受到国内外学者的深度重视则是在 2014 年该算法在机器翻译领域获得巨大成功后[62]。从此，研究者将其广泛应用于金融预测方面并取得了突破性进展[63]。Di Persio 和 Honchar（2017）在研究谷歌公司股票价格的走向时分别使用了 RNN、LSTM、GRU 模型，发现 LSTM 模型预测效果最佳、预测性能稳定[64]。陈卫华（2018）采用 5 分钟高频交易数据计算出日收益率和已实现波动率，利用 LSTM 模型对股票波动率进行预测，并以 RMSE、MAPE、QLIKE 等 6 项作为评价指标，证明 LSTM 相较于其他 19 种模型的预测能力最好[25]。杨青和王晨蔚（2019）以股票历史价格数据为输入特征，对可以反映市场整体变动状况的全球 30 个股票指数应用 LSTM 模型，发现预测期限无论是在短期、中期还是长期，平均预测精度和稳定度均优于支持向量回归（Support Vector Regression，SVR）、多层感知机（Multilayer Perceptron，MLP）以及 ARIMA 模型[65]。Kavinnilaa 等（2021）在文中指出 LSTM 模型通过长期依赖学习数据特征，加之有效的更新、