改进的 NSGA-III-XGBoost 算法在股票预测中的应用

何泳[1]，李环[1]

1.东莞理工学院 计算机科学与技术学院，广东 东莞 523000

摘　要：为提高股票预测的准确度和减少运行时间,提出了一种改进的非支配排序遗传算法与极致梯度提升树模型相结合（INSGA-III-XGBoost）的股票预测模型。该模型特征工程包括小波分解、扩展特征、数据清洗、归一化。模型采用两种过滤式特征选择的集成信息初始化种群优化 NSGA-III 算法，以最大化准确度和最小化解的解决方案大小作为优化方向，使用多染色体混合编码的方式同步进行特征选择和优化模型参数。最后，将选择的特征子集和参数输入 XGBoost 训练预测并迭代优化。实验结果表明，INSGA-III-XGBoost 算法与未改进的多目标特征选择算法和单目标特征选择算法相比，平均准确度最高、解方案最小、运行时间最短；与深度学习模型相比，不仅准确度更高、运行用时大幅减少，并且该模型具有可解释性。

关键词：多目标优化；特征工程；特征选择；股票预测

文献标识码：A　　中图分类号:TP399

Application of Improved NSGA-III-XGBoost Algorithm in Stock Forecasting

Yong He[1], Huan Li[1]

1.College of Computer Science and Technology, Dongguan University of Technology, Dongguan Guangdong 523000, China

Abstract: To improve the accuracy of stock forecasting and reduce the running time, a stock forecasting model combining an improved non-dominated sorting genetic algorithm and extreme gradient boosting tree model (INSGA-III-XGBoost) is proposed. The model feature engineering includes wavelet decomposition, extended features, data cleaning, and normalization. The model uses the integrated information of two types of filtered feature selection to initialize the population optimization NSGA-III algorithm, maximize the accuracy and minimize the solution size of the solution as the optimization direction, and use the multi-chromosome hybrid encoding method to simultaneously perform feature selection and optimize model parameters. The selected feature subsets and parameters are input into XGBoost for training and forecasting, and iteratively optimizes according to the evaluation metrics. The experimental results show that compared with the unimproved multi-objective feature selection algorithm and single-objective feature selection algorithm, the INSGA-III-XGBoost algorithm has the highest average accuracy, the smallest solution scheme, and the shortest running time; compared with the deep learning model, it not only has higher accuracy, but the runtime is greatly reduced as well, and the model is interpretable.

Keywords: multi-objective optimization; feature engineering; feature selection; stock forecasting

股票预测是计算机科学与金融交叉的经典问题。由于近年来人工智能技术的不断发展和股票数据易获得的特点，越来越多的模型用于预测股票[1]。由于机器学习模型具有更强大的大数据处理能力和学习能力，能够处理输入特征和预测目标之间的

非线性关系，因此其预测能力通常比传统的基本面分析的方式的更强[2]。通过准确的股票价格方向变动预测，投资者可以把握买卖时机，从而战胜市场并获取利润[3]。

目前的研究中，基本的股票数据处理流程为：数据预处理、特征工程、模型训练、优化、预测和评估。然而，大部分的工作都集中在预测算法而忽视了特征工程。即使深度学习可以做到全自动的特征工程，也需要在输入模型之前进行数据预处理，好的特征工程可以使预测模型达到更好的性能的同时减少运行资源[4]。因此，本文首先在特征工程方面进行数据降噪和生成技术指标，而后使用结合了多目标优化（NSGA-III）算法和机器学习算法（XGBoost）的股票预测模型进行特征选择并对股票的变动方向进行预测。

本文股票预测模型的优点如下：1.高效。本文提出的算法与深度学习中长短时记忆神经网络（Long short term memory neural network，LSTM）神经网络相比，在相同的优化次数下，准确度比后者高的同时，所需运行时间不到后者的 1%。2.可解释性。深度学习模型普遍存在“黑盒”问题[5]，无法对特征的重要性进行评估，而本文提出算法可以得出重要性最高的特征，以供后续研究。3.高准确度和稳定性，本文算法与其他基准研究相比，体现了其预测能力和应对不同市场数据的预测稳定性。

本文后续内容安排如下：第一部分介绍相关工作、第二部分介绍模型与方法、第三部分是实验与分析，最后在第四部分总结全文。

1 相关工作

近几年，深度学习在图像识别、语音识别和自然语言处理等领域非常热门，其中也包括时间序列分析。在时间序列分析中， LSTM 由于其特殊的门结构，可以记忆过去一段长度的输入并解决了循环神经网络（Recurrent neural network，RNN）中的梯度消失问题，成为研究的热门。其中文献[6]使用遗传算法将均方根误差（RMSE）作为适应度函数，优化了神经网络。文献[7]采用数据增强的方式，并用一个预测 LSTM 和一个防止过拟合 LSTM 来提高预测性能。文献[8]提出一个复合模型，LSTM 结合了经验小波分解和异常值鲁棒极端学习机。除了深度学习多种机器学习模型也用来开发股票预测系统。例如文献[9]使用基于树的集成学习方式，文献[2]使用遗传算法优化的 XGBoost，文献[10]采用非线性高斯核函数的权重支持向量机（support vector machine, SVM）进行特征工程，并用权重 K 邻近算法预测价格。

特征工程是 AI 技术的重要组成，许多研究在这一部分改进来提高模型的预测性能。文献[11]、文献[6]和文献[9]在原始的历史数据上生成技术指标来预测价格。而文献[9]在拓展特征后，采用特征提取的方式获得新的指标。然而，过多的特征输入不一定能提高模型的性能，反而可能导致“维度诅咒”的问题，造成不必要的计算消耗和模型预测能力的下降[12]。

基于以上研究的启发，本文提出的算法在拓展特征后，采用特征选择的方式，去除不相关的和冗余的特征，减少不必要的计算开销的同时提高模型的预测性能。进化算法通过启发式搜索策略获得最佳特征子集，因为其高效的全局搜索方式被广泛应用于特征选择问题。目前大多数研究采用单目标的方式优化分类精度或分类误差来解决特征选择的问题，而该问题可以作为最大化预测性能和最小化特征数量的多目标问题。在实际应用中，如果能选择较小的解集并保持较高的预测性能，那么就能减少计算量的同时提高预测性能。因

此，本文的预测系统可以看成多目标特征选择问题，用多目标算法解决。

2 方法与模型

为了解决高维数据的处理问题，本文通过特征选择方式移除与目标相关性低的特征和冗余的特征，从而提高计算效率和模型性能。本文提出 INSGA-III-XGBoost 算法使用多目标算法进行同步特征选择和参数优化，选择的特征子集和参数输入 XGBoost 模型进行训练预测。在本文的研究中将选择的特征数和分类的准确度作为两个目标。其中，选择特征数量的目标函数表示为：

$$f_1(Z) = \frac{1}{D}\sum_{j=1}^{D} z_j \qquad (1)$$

分类准确度的计算公式为：

$$f_2(Z) = \left(\frac{1}{l}\sum_{l=1}^{n} \frac{N_{Car}}{N_{All}}\right) \times 100\% \qquad (2)$$

式中，Z为解码方案，D 为维度（特征的个数），N_{Car}为正确预测的样本数，N_{All}是所有样本的数量。

如图 1 所示，本文采用多染色体混合编码的方式，第一条染色体编码了所有特征，染色体的长度等于数据的特征数量，其中 0 代表未选择该特征，1 选择该特征将被保留。第二条染色体编码了 XGBoost 的关键参数。图 3 中的 E 为 XGBoost 树的数量，F 为 XGBoost 最大特征数，L 为 XGBoost 学习率。

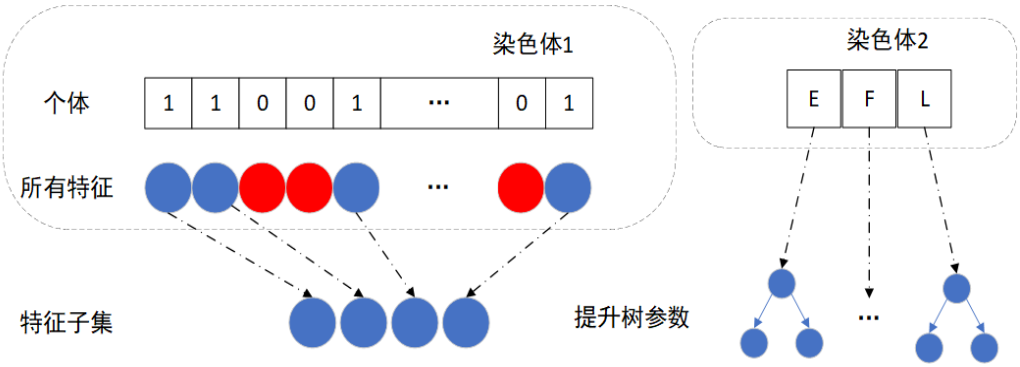


图 1 个体表示

Fig.1 Presentation of individual

卡方检验方法(Chi2)可以对特征进行相关性分析，因此在 NSGA-III 算法的初始化阶段，可以用 Chi2 先评估特征，在所有的特征变量里提取出与目标更相关的特征，并减少相关度低的冗余特征，用于初始化种群以提高性能。公式如下：

$$x = \frac{(实际值 - 理论值)^2}{理论值} \qquad (3)$$

其*实际值*指变量 x 的实际频数，*理论值*指假设变量 x 与目标变量 y 之间独立时，x 的理论频数。先通过 Chi2 评估特征,获得每个特征变量对目标变量的卡方值，然后根据大小排序，选择排名靠前的特征，即是与目标变量 y 更相关的特征[13]。通常在 Chi2 算法初始化过程中，需要保留大部分的评分较高的特征和不相关的特征的小部分以保持初始化的多样性（考虑特征之间的相互作用）。因此，通过实验对比，本文选择 80%最有用的特征。而如何从选择的特征中得到最合适的特征组合，并且确定合适的特征数量又是需要考虑的问题。在本文中，采用混合初始化[14]的方法解决该问题：首先基于卡方值排序，从初始特征中选择得分最高的 80%的特征保存在WR中，对于所有个体中的 80%的个体,若个体选择的特征在WR且初始矩阵XG中选择了该特征(对应的值为 1)，则保留该特征。对于所有个体中的 20%的个体，若个体选择的特征不在WR且初始矩阵XG选择了该特

征，则保留该特征。

算法 1 混合初始化

种群大小ps，特征数D，初始矩阵XG，前 80%特征矩阵WR，记录矩阵PF

```
for i ≤ ps do
    for 80%的个体 do
        for j ≤ D do
            if (XG_ij == 1) and (j ∈ WR) then
                PF_ij = 1;
            else
                PF_ij = 0;
            end
    end
    for 20%的个体 do
        for j ≤ D do
            if (XG_ij == 0) and (j ∈ WR) then
                PF_ij = 1;
            else
                PF_ij = 0;
            end
    end
```

XGBoost 创建后每个属性的重要性得分可以直接获得，该得分衡量特征在提升树构建时的重要程度。在单个决策树，每颗树根据特征对性能度量改进的量计算属性重要性。在提升树中，单个特征对性能改进的程度越大，权值越大，并将被更多提升树所选择，重要性越高。最后根据属性在所有提升树中的重要性加权求和并平均，得到最终的重要性评分。由于其与 Chi2 的评估方法不同，因此会得到不同的评估结果。基于 Chi2 和 XGBoost 的集成学习种群初始化过程如下：第一部分种群使用 Chi2 评估特征的卡方值，根据其大小将特征从大到小排列，然后由算法 1 得到初始种群 a，再将种群 a 作为 XGBoost 的输入，使用 TPE(Tree-structured Parzen Estimator)算法[15]优化 50 次参数，得到原种群和准确度合并的新种群 A。第二部分种群根据 XGBoost 评估特征的重要性得分，根据重要性得分将特征从大到小排列，再由算法 1 得到初始种群 b,再将种群 b 作为 XGBoost 的输入，并使用 TPE 算法优化 50 次参数，得到原种群和准确度合并的新种群 B。最后将新种群 A 和 B 合并，根据准确度排序，选择前 50%的个体作为最终的初始化种群 P。

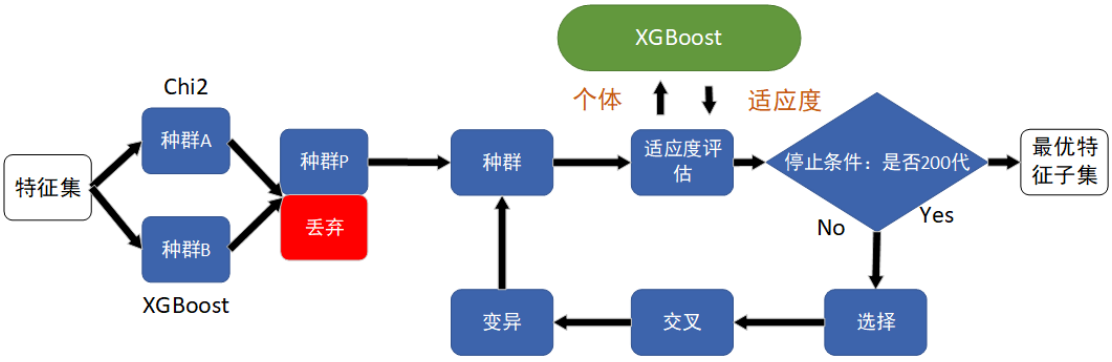


图 2 INSGA-III-XGBoost 算法

Fig.2 INSGA-III-XGBoost Algorithm

INSGA-III-XGBoost 算法通过收集并更新每代中达到的最高准确度的解集信息，搜索近似最优或最优解。图 2 展示了上述 INSGA-III-XGBoost 的过程。NSGA-III 算法中某一代种群中所有个体将 XGBoost 训练评估的适应度，经过交叉、变异找到最优的解。XGBoost 全称极端梯度提升树，它是在数据科学竞赛中占据主导地位的非深度学习算法。XGBoost 由陈天奇博士[16]设计开发，模型设计只关注性能和效率，

能够并行的将多个弱分类器（决策树）通过结果加权的方式合成强分类器（提升树），是工程领域最好用的算法之一。

3 实验结果与讨论

本文所有的实验均在如下配置的计算机中运行。硬件信息：英特尔 i5-9500（3.00GHZ）处理器、8GB RAM；软件信息：Python 3.8.5、Visual Studio Code 1.67.1、Jupyter notebook 6.4.6。因为市场状态可能潜在地影响股票预测的效果，因此从不同发展程度的市场选择指数有助于解释算法的鲁棒性。本文选择的 3 只市场指数，道琼斯指数代表最发达市场指数，恒生指数代表比较发达市场的指数，上证 300 代表发展中市场的指数，所有数据均通过 Investing.com 下载。数据样本的时间段为：2008 年 7 月 1 日至 2016 年 9 月 30 日。

3.1 数据降噪

小波变化具有处理不平稳的金融时间序列的能力，因此本文中使用了小波变化进行数据降噪。小波变换的关键特性是与傅里叶变换相比，它可以同时分析金融时间序列的频率分量。因此它可以有效地处理高度不规则地金融时间序列[17]。本文使用三层 sys8 小波将指数价格序列分解为时域和频域。

3.2 生成技术指标

本文将建立两个指标集。一个是在前人的研究中常用的指标，一个是本文生成的指标，两个指标集进行对比。表 1 展示了前人研究中常用技术指标。

表 1 技术指标集

Table 1 Technical indicator set

分类	指标
历史数据	最高价、最低价、开盘价、收盘价、成交量
技术指标	异同移动平均线 (MACD)、顺势指标 (CCI)、均幅指标（ATR）、布林线（BOLL）、20 日指数移动平均值（EMA20）、5/10 日移动平均（MA5/MA10）、6/12 月动力指标（MTM6/MTM12）、变动率（ROC）、随机动量（SMI）威廉变异离散量（WVAD）
经济变量	汇率、利率

原始的历史数据只包括开盘价、收盘价、最高价、最低价和成交量。本文通过生成技术指标的方式，将初始的 5 维数据拓展为 81 维数据。所有的技术指标均通过 TA-Lib 库生成，可以分为六组，分别是重叠指标、动量指标、成交量指标、波动率指标、价格转换指标和循环指标。

3.3 数据清洗和归一化

将 1950 个交易日的数据集划分，其中训练集 85%，测试集 15%。训练集分为前 82%训练模型，后 18%用来验证模型。公式（4）将数据集映射到[0,1]之间进行归一化。

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

3.4 性能衡量指标

实验采用下列的常用分类指标衡量算法性能。

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{6}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{7}$$

$$F1-\text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

$$AUC = \int_0^1 \frac{TP}{TP+FN} d\frac{FP}{FP+TN} \tag{9}$$

其中 TP 为真正率，TN 为真负率，FP 为假正率，FN 为假负率。

3.5 单目标、多目标、改进多目标对比

3.5.1 分类指标

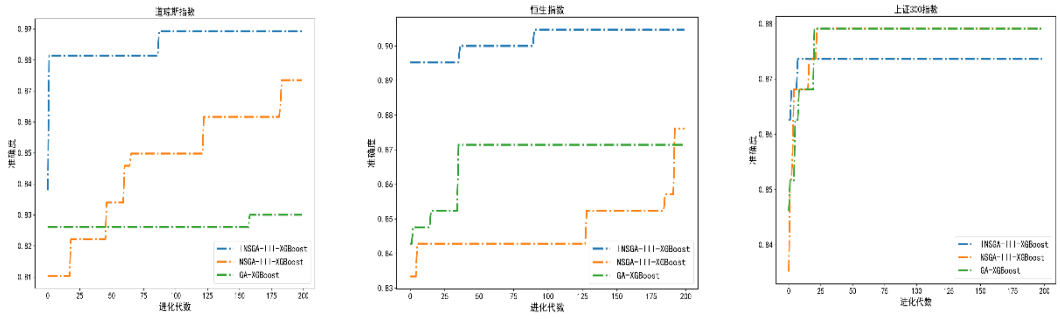
分别采用基于单目标的方法 GA 和基于多目标的方法 NSGA-III,以及基于改进后的多目标的方法 INSGA-III 结合 XGBoost 对三个数据集进行实验，比较不同特征选择算法的性能。采用的参数设置如下：进化代数 200 代，种群大小 20，个体染色体数 2，交叉率 1，变异率 0.05。实验数据如表 3 所示。表中可以看出在三个

数据集中，INSGA-III-XGBoost 算法在准确度、F1-score、AUC 上均取到 2 个最佳，两个多目标算法选择的特征数均比单目标算法选择的特征较少。实验结果表明，从分类指标评价的角度上看，单纯把单目标特征选择问题转换为多目标特征选择问题效果不一定会更好，而本文改进 INSGA-III 算法则提升了多目标算法的效果，总体表现优于未改进的多目标算法和单目标算法。不同算法在进化过程中的准确度变化如图 3（a、b、c）所示。

表 2 分类指标对比

Table 2 Comparison of classification metrics

		Accuracy (%)	F1-score (%)	AUC	选择特征数
INSGA-III-XGBoost	道琼斯指数	**88.93**	**90.54**	**88.60**	22
	恒生指数	**90.48**	**92.31**	**89.72**	11
	上证 300	86.81	85.19	86.65	26
NSGA-III-XGBoost	道琼斯指数	87.35	89.33	86.70	11
	恒生指数	87.62	90.15	86.25	4
	上证 300	**87.91**	**86.59**	**87.89**	43
GA-XGBoost	道琼斯指数	85.77	88.31	84.52	50
	恒生指数	88.57	90.84	87.48	43
	上证 300	**87.91**	86.25	87.64	41



（a）道琼斯指数准确度变化
(a) Accuracy changes in DJIA dataset

（b）恒生指数准确度变化
(b) Accuracy changes in HangSeng dataset

（c）上证 300 指数准确度变化
(c) Accuracy changes in CSI300 dataset

图 3 进化过程准确度变化

Fig.3 Change of Accuracy in the process of evolution

3.5.2 运行时间对比

表 3 展示了三种算法的运行时间对比。表中可以看出，本文提出的 INSGA-III-XGBoost 算法所需的运行时间最少，相比 NSGAIII-XGBoost 算法平均运行时间缩短了 39.4%，而相比于采用单目标优化的 GA-XGBoost 算法，平均时间缩短了 83.28%。INSGA-III-XGBoost 算法的运行时间方差最小，体现了 INSGA-III-XGBoost 算法在运行时间方面的稳定性。其运行时间的较小的原因是，算法在选择了较少的特征的同时选择了合适的提升树结构并动态调整了学习率，避免了大量不必要的计算开销，从而提高了运行效率。

表 3 运行时间对比

Table 3 Comparison of processing time

	最大值 (s)	最小值 (s)	平均值 (s)	方差 (s)
INSGA-III-XGBoost	**227.60**	**252.22**	**262.90**	**13.16**
NSGA-III-XGBoost	462.26	400.77	433.76	31.07
GA-XGBoost	1778.05	1370.13	1572.81	203.97

3.5.3 综合比较

运行时间、选择的特征数、准确度是三个最重要的指标。表 4 展示了三个数据集的平均指标的对比。表中可以看出，本文提出的 INSGA-III-XGBoost 算法综合运行时间最短，选择的特征数最少，且准确度最高，即性能表现最好。

表 4 平均指标综合比较

Table 4 Comprehensive comparison of average metrics

	运行时间 (m)	特征数	Accuracy (%)
INSGA-III-XGBoost	**4.38**	**20**	**88.74**
NSGA-III-XGBoost	7.23	20	87.63
GA-XGBoost	26.21	45	87.42

3.6 不同特征子集对比

以恒生数据集为例，对 4 种不同的特征数据集进行对比，4 种数据集分别输入 XGBoost 训练预测，结果如表 5 所示。其中，历史特征数据集，仅包含原始的五个特征。其他特征数据集包含前人研究中常用的 18 个特征。所有特征数据集是本实验拓展特征阶段得到的 81 个特征，最优子集数据集为本文提出算法 INSGA-III-XGBoost 选择出的最佳特征子集，相比于所有特征数据集，减少了 70 个特征。

表 5 恒生数据集的特征子集对比

Table 5 Comparison of feature subset in HangSeng dataset

	特征数	Accuracy (%)	F1-score (%)	AUC
历史特征数据集	5	77.62	82.53	75.12
所有特征数据集	81	81.43	85.71	78.68
其他特征数据集	18	78.51	83.12	80.70
最优子集数据集	11	**90.48**	**92.31**	**89.72**

INSGA-III-XGBoost 算法从 80 个特征数据集中选择 11 个最佳特征子集，对这 11 个特征进行分析，这是神经网络的“黑盒”模型不具备的优势。图 4 展示了该最佳特征子集中的特征，按照重要性得分降序。BOLLM 即布林线的中线是所有特征中重要性得分最高的，这说明其对预测股票的走势作用最大。

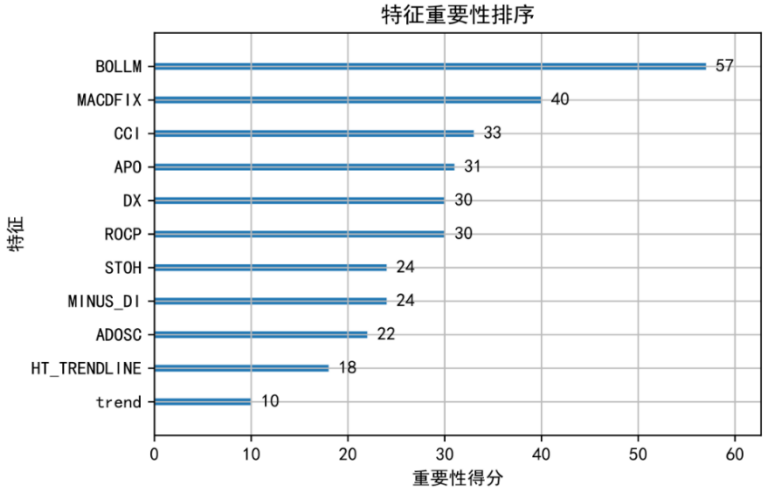


图 4 最佳子集特征重要性得分

Fig.4 Importance score of features of optimal subset

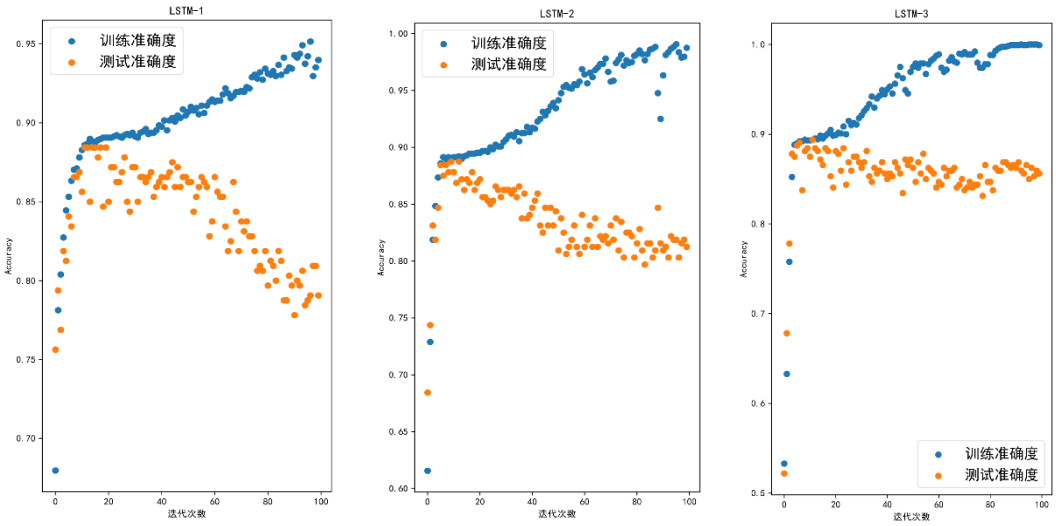
3.7 与基准模型对比

本算法与经典的机器学习模型和不同层数的深度学习模型 LSTM，和双向 LSTM（Bidirectional long short term memory, BiLSTM）比较。以恒生数据集为例，表 6 看出本文的算法具有最高的准确度、F1-score 和 AUC。三种不同层数的 LSTM 迭代 100 次的实验结果如图 5（a、b、c）所示，LSTM 神经网络的训练准确度提高的很快，但是验证准确度先增后减，造成了过拟合的问题。

表 6 基础模型对比

Table 6 Compare with base model

	Accuracy (%)	F1-score (%)	AUC
INSGA-III-XGBoost	**90.48**	**92.31**	**89.72**
SVM	86.67	89.47	85.01
K 邻近	84.29	87.82	81.93
XGBoost	81.43	85.71	78.68
随机森林	67.62	74.44	64.92
决策树	58.57	63.29	58.70
极限树	63.81	68.07	64.11
LSTM layer1	83.91	83.04	83.24
LSTM layer2	86.97	86.37	86.32
LSTM layer3	86.97	86.41	86.41
BiLSTM layer1	81.99	80.84	81.36
BiLSTM layer2	89.66	89.08	88.78
BiLSTM layer3	87.74	86.87	86.42



（a）1 层 LSTM 训练和测试准确度

(a) Training and testing accuracy of 1-layer LSTM

（b）2 层 LSTM 训练和测试准确度

(b) Training and testing accuracy of 2-layer LSTM

（c）3 层 LSTM 训练和测试准确度

(c) Training and testing accuracy of 3-layer LSTM

图 5 不同层数 LSTM 迭代过程准确度变化

Fig.5 Change of accuracy in iterative process of LSTM with different layers

深度学习因为其强大的预测能力而应用于时间序列预测，但是模型的能力在很大程度上依赖于神经网络结构和超参数的调整。本文实验使用 TPE 算法迭代优化 LSTM 神经网络结构 200 次。表 7 展示了 INSGA-III-XGBoost 算法和 TPE-LSTM 算法的运行时间和准确度。表中可以看出，两者的准确度相差较小，但是本文提出的算法的运行时间仅为 TPE-LSTM 的 0.99%。

表 7 TPE-LSTM 对比

Table 7 Compare with TPE-LSTM

	运行时间（m）	Accuracy（%）
TPE-LSTM	442.16	89.47
INSGA-III-XGBoost	**4.38**	**90.48**

3.8 与基准研究对比

表 8 为本文与近几年来的基准研究对比，对比结果验证了本文提出模型的优越性。三大市场平均准确度比其他基准研究的准确度更高，从而验证了本文提出模型适应不同市场数据的能力。

表 8 与基准研究对比

Table 8 Compare with benchmark studies

数据集	作者	方法	Accuracy	F1-score	AUC
Apple	(Jin, et al., 2020) [18]	S_EMDAM+LSTM	70.56	N/A	N/A
S&P 500 index	(Yang et al., 2020) [19]	CNN + LSTM	63.16	62.50	N/A
CITIC security	(Long et al., 2020) [20]	CNN + BiLSTM	75.89	N/A	73.10
Tokyo Stock Exchange index	(Qiu & Song, 2016) [21]	GA-ANN	81.27	N/A	N/A
KOSPI index	(Chung & Shin, 2020) [22]	GA-CNN	73.74	N/A	N/A
NASDAQ index	(Singh et al., 2017) [23]	(2D)2 PCA+ DNN	68.21	N/A	N/A
Bank of America	(Ampomah et al., 2020) [24]	集成学习	84.63	85.05	**92.80**
恒生指数	本文	INSGA-III-XGBoost	**90.48**	**92.31**	89.72

三大市场平均	本文	INSGA-III-XGBoost	88.74	89.35	88.32

3.9 其他多目标算法对比

将本文的提出的 INSGA-III 算法与其他多目标优化算法分别结合 XGBoost 比较性能，表 9 展示了恒生数据集实验结果。INSGA-III-XGBoost 算法的准确度最高，达到 90.48%。图 6 为各多目标算法得到的帕累托前沿，纵轴为分类误差，即 1-准确度。相比于其他多目标算法，INSGA-III 优化效果最好，具有较小的解决方案大小和较低的分类错误率。

表 9 多目标算法结合 XGBoost 对比

Table 9 Compare with multi-objective algorithm combined with XGBoost

	非支配个体数	Accuracy (%)	F1-score (%)	AUC	选择特征数
INSGA-III-XGBoost	3	**90.48**	**92.31**	**89.42**	11
NSGA-III-XGBoost	4	87.62	90.15	86.25	4
NSGA-II-XGBoost	7	87.62	90.15	86.25	4
AWGA-XGBoost	4	88.57	90.98	87.02	8
RVEA-XGBoost	3	89.05	91.25	87.87	9

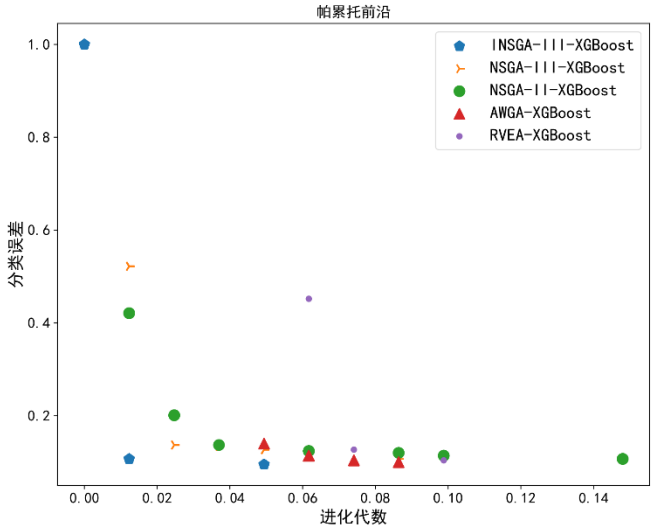


图 6 多目标算法帕累托前沿

Fig.6 Pareto Front of multi-objective algorithm

4 结语

本文提出的 INSGA-III-XGBoost 算法通过将两种过滤式特征选择集成的方法初始化种群，并将股票预测问题作为多目标问题，以最大化准确度和最小化解的解决方案大小作为优化方向,采用多染色体混合编码的方式同步优化了特征选择和 XGBoost 参数，对比其他基准研究具有最快的处理速度，解方案最小，预测准确度最高。在特征工程方面首先生成 81 个特征，将原始历史数据 5 个特征的预测准确度从 77.62%提升到 81.43%，而特征选择再选择其中 11 个特征作为最优子集预测准确度提升到 90.48%，克服了“维度诅咒”的问题。在所有基础的机器学习模型和深度学习模型中本算法的预测性能最好。对比原始多目标算法和单目标算法运行时间分别缩短了 39.4%和 83.28%，运行效率高，适合短期交易系统的预测需求。对比默认参数的深度学习算法，预测性能平均高 3%，而对比经过 200 代

参数优化的 TPE-LSTM 虽然准确度只高 1%，但是运行时间仅为它的 0.99%，并且本文模型具有可解释性，实验结果给出了预测恒生指数走向的前 11 个关键特征及其重要性得分。

参考文献

[1] Jiang W. Applications of deep learning in stock market prediction: recent progress[J]. Expert Systems with Applications, 2021, 184: 115537.

[2] Ding G, Qin L. Study on the prediction of stock price based on the associated network model of LSTM[J]. International Journal of Machine Learning and Cybernetics, 2020, 11(6): 1307-1317.

[3] Yun K K, Yoon S W, Won D. Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process[J]. Expert Systems with Applications, 2021, 186: 115716.

[4] Chollet F. Deep learning with Python[M]. Simon and Schuster, 2021: 98.

[5] Montavon G, Samek W, Müller K R. Methods for interpreting and understanding deep neural networks[J]. Digital Signal Processing, 2018, 73: 1-15.

[6] Chung H, Shin K. Genetic algorithm-optimized long short-term memory network for stock market prediction[J]. Sustainability, 2018, 10(10): 3765.

[7] Baek Y, Kim H Y. ModAugNet: A new forecasting framework for stock market index value with an overfitting prevention LSTM module and a prediction LSTM module[J]. Expert Systems with Applications, 2018, 113: 457-480.

[8] Liu H, Long Z. An improved deep learning model for predicting stock market price time series[J]. Digital Signal Processing, 2020, 102: 102741.

[9] Ampomah E K, Qin Z, Nyame G. Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement[J]. Information, 2020, 11(6): 332.

[10] Chen Y, Hao Y. A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction[J]. Expert Systems with Applications, 2017, 80: 340-355.

[11] Naik N, Mohan B R. Stock price movements classification using machine and deep learning techniques-the case study of indian stock market[C]//International Conference on Engineering Applications of Neural Networks. Springer, Cham, 2019: 445-452.

[12] Li J, Cheng K, Wang S, et al. Feature selection: A data perspective[J]. ACM computing surveys (CSUR), 2017, 50(6): 1-45.

[13] Kumar B S, Ravi V, Miglani R. Predicting Indian stock market using the psycho-linguistic features of financial news[J]. Annals of Data Science, 2021, 8(3): 517-558.

[14] Xue Y, Xue B, Zhang M. Self-adaptive particle swarm optimization for large-scale feature selection in classification[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2019, 13(5): 1-27.

[15] Hyperopt [CP/OL]. [2021-12-31] https://github.com/hyperopt.

[16] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.

[17] Bao W, Yue J, Rao Y. A deep learning framework for financial time series using stacked autoencoders and long-short term memory[J]. PloS one, 2017, 12(7): e0180944.

[18] Jin Z, Yang Y, Liu Y. Stock closing price prediction based on sentiment analysis and LSTM[J]. Neural Computing and Applications, 2020, 32(13): 9713-9729.

[19] Yang C, Zhai J, Tao G. Deep learning for price movement prediction using convolutional neural network and long short-term memory[J]. Mathematical Problems in Engineering, 2020, 2020(6):1-13.

[20] Long J, Chen Z, He W, et al. An integrated framework of deep learning and knowledge graph for prediction of stock price trend: An application in Chinese stock exchange market[J]. Applied Soft Computing, 2020, 91:

106205.

[21] Qiu M, Song Y. Predicting the direction of stock market index movement using an optimized artificial neural network model[J]. PloS one, 2016, 11(5): e0155133.

[22] Chung H, Shin K. Genetic algorithm-optimized multi-channel convolutional neural network for stock market prediction[J]. Neural Computing and Applications, 2020, 32(12): 7897-7914.

[23] Singh R, Srivastava S. Stock prediction using deep learning[J]. Multimedia Tools and Applications, 2017, 76(18): 18569-18584.

[24] Ampomah, E. K., Qin, Z., & Nyame, G. (2020). Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement. Information, 11(6): 332.

基金项目：高等学校“创新强校工程“创新项目：基于机器学习算法的科技成果转化精准推荐模型研究（2018KTSCX222）

作者简介：何泳(1998-)，男，硕士研究生，研究方向：智能优化、机器学习,E-mail:987103622@qq.com；

李环(1977-)，女，博士，副教授，研究方向：多媒体信息网络安全、人工智能 E-mail：lihuan@dgut.edu.cn；

本文有任何疑问请联系作者：何泳，手机号是：15707536251，邮箱是：987103622@qq.com，通讯地址是：广东省东莞市松山湖管委会大学路一号东莞理工学院松山湖校区 24 号楼。