基于过滤式改进的多目标特征选择算法在股票预测中的应用

何泳

（东莞理工学院 计算机科学与技术学院，广东 东莞 523000）

摘 要:

关键词: 多目标优化；特征工程；特征选择；股票预测

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，因为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（NSGA3）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在时间序列分析中最热门的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.可解释性好。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。本算法与单目标和其他多目标算法相比，体现了其适应不同市场数据的预测能力和稳定性。

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，长短时记忆神经网络（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]提出一个复合预测模型EWT-LSTM-ORELM。其流程为先用经验小波分解（EWT），再用LSTM结合粒子群优化算法，最后再用异常值鲁棒极限学习机（ORELM）后处理。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SVM进行特征工程，并用权重KNN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，该方法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。

1. 方法与模型
   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[13]。以恒生指数为例，本文使用三层sys8小波将指数价格序列分解为时域和频域，图1为降噪前的序列，为实验所用数据中恒生指数的收盘价变化而经过三层小波分解后的收盘价如图2所示。

图形用户界面, 图表, 散点图

描述已自动生成

图 1 恒生指数收盘价

图形用户界面, 图表, 散点图

描述已自动生成

图 2 三层小波分解后

* 1. 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 技术指标集一

|  |  |
| --- | --- |
| 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（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。

* 1. 数据清洗和归一化

从2008年7月1日到2016年9月30日到2038个交易日内删除由于拓展特征造成的前88个空值，剩1950个交易日作为数据集输入。将所有数据分为训练集和测试集，其中训练集占85%（前1650交易日），测试集占15%（后300交易日）。测试集用来评估最终选出特征子集。训练集分为前82%（1350交易日）训练模型，后18%（300交易日）用来验证模型。所有的值都由公式（1）映射到[0,1]之间。

（1）

* 1. 改进的多目标特征选择算法

为了解决高维数据的处理问题，通常采用特征提取和特征选择的方法对数据进行降维。特征提取通过原始特征的混合产生新的特征，因此失去了原始特征的解释性和可理解性。特征选择通过移除与目标相关性的底的特征和冗余的特征提高了计算效率和模型性能[12]。本文提出INSGA3-XGBoost算法基于过滤特征选择的方式，通过移除不相关的冗余特征提高分类器性能，在NSGA3种群初始化阶段结合两种过滤方法集成的过滤式方法，进行选择特征和种群的前置处理，进一步提高了算法的效率。

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：

（2）

分类准确度的计算公式为：

（3）

式中，为解码方案，D为维度（特征的个数），为正确预测的样本数，是所有样本的数量。在本文中使用改进的NSGA3算法同步进行了特征选择和分类器结构优化，加快了算法的收敛速度，缩短了运行时间，并且提升了算法性能。

如图3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2中的E为XGBoost树的数量，F为XGBoost最大特征数，L为XGBoost学习率。算法根据选择的特征子集和关键参数，训练对应的XGBoost模型，计算得到选择的特征数和准确度将被用来评估优化，并且不断产生更好地解集，推动种群中的个体寻找到全局最优。

由于卡方检验方法(Chi2)可以对分类变量进行相关性分析，因此在NSGA3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，并减少冗余特征，用来初始化种群以提高性能。公式如下：

（4）

其中实际值指变量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[14]。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。如何从选择的特征中得到合适的特征，并且选择合适数量的特征又是需要考虑的问题。在本文中，采用混合初始化[15]的方法解决该问题：首先基于卡方值排序，从初始特征中选择卡方值最高的80%的特征保存在中，对于所有个体中的80%的个体,若个体选择的特征在且初始矩阵中选择了该特征(对应的值为1)，则保留，否则不保留该特征。对于所有个体中的20%的个体，若个体选择的特征不在且初始矩阵选择了该特征(对应的值为1)，否则不保留该特征。

|  |
| --- |
| 算法1：混合初始化  种群大小，特征数，初始矩阵，前80%特征矩阵，记录矩阵 |
| 4. ; 6. ;     12. ; 14. ; |

|  |
| --- |
| 算法2：集成学习种群初始化 |
| 1. chi2评估特征的卡方值，根据卡方值大小将特征从大到小排列。 2. 由算法1得到初始种群a。 3. 种群1作为 XGBoost的输入，并使用TPE算法[16]优化50次参数，得到原种群和准确度合并的新种群A。 4. 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 5. 由算法1得到初始种群b。 6. 种群2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。 7. 将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。 |

XGBoost的提升树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，将两部分得到的启发信息集成得到的信息更具有适用性。算法2显示了基于Chi2和XGBoost的集成学习种群初始化过程。

INSGA3-XGBoost算法通过每代中达到的最高准确度的解集信息，收集并更新近似最优或最优解。图4展示了上述INSGA3-XGBoost的过程。初始的种群有20个个体，个体采用多染色体编码，每个个体有与总输入特征的数量相同长度的染色体和控制XGBoost三个关键参数的染色体。NSGA3算法中某一代种群中所有个体由XGBoost训练评估的准确度和当前个体所选择的特征数量作为多目标算法优化的两个目标，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[17]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。

1. 实验结果

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。3个指数历史数据包括开盘价、收盘价、最高价、最低价、成交量，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。

本实验采用的参数设置如表2所示，其中n代表种群大小。本文采用下列的常用分类指标衡量分类器的性能。

（5）

（6）

（7）

（8）

（9）

其中TP为真正率，TN为真负率，FP为假正率，FN为假负率。

表 2参数设置

|  |  |
| --- | --- |
| 参数类型 | 参数设置 |
| 进化代数  种群大小  个体染色体数 | 200  20  2 |
| 交叉率  变异率 | 1  1/n |

* 1. 单目标、多目标、改进多目标对比

3.2.1 分类指标对比

3.2.2运行时间对比

* 1. 不同特征子集对比
  2. 与基准模型对比
  3. 与基准研究对比
  4. 与深度学习对比
  5. 其他多目标算法对比

1. 结语