人工智能算法在定量金融领域的应用探索

Huang YuHang
SUSTech CSE
12011129@mail.sustech.edu.cn

Luo Jiajun
SUSTech CSE
12012023@mail.sustech.edu.cn

Zhao Ketai
SUSTech CSE
12110306@mail.sustech.edu.cn

Liao Zetong
SUSTech CSE
12011417@mail.sustech.edu.cn

Li Ming
SUSTech Mathematics
12110435@mail.sustech.edu.cn

January 7, 2024

1. 摘要

量化金融作为金融领域的一支新兴分支，通过运用数学模型和统计分析，试图发现市场中的规律性，并基于这些规律性制定交易策略。在过去几十年的发展中，量化金融已经成为金融机构和个人投资者的主要手段之一，而其逐渐崭露头角的背后离不开人工智能技术的不断创新与应用。我们梳理了量化金融的发展背景和历史沿革，探讨了人工智能在该领域中的关键作用。重点对卷积神经网络、随机森林、配对算法、与遗传算法在量化金融中的应用进行了研究与实验。其中，卷积神经网络以其强大的特征提取能力在量化模型中展现出色，随机森林则通过集成学习有效地处理复杂的市场数据。配对算法的应用使得投资组合管理更为精准，而遗传算法在参数优化中展现出强大的搜索和优化能力。这些人工智能技术的整合为量化金融提供了新的解决方案，提高了交易策略的稳定性和效益。未来，随着人工智能技术的不断发展，量化金融领域将迎来更多创新与突破，为投资者提供更可靠的决策支持。

关键词：量化金融、人工智能、卷积神经网络、随机森林、配对算法、遗传算法

2. 背景

2.1. 量化金融及其历史发展

量化金融是一种利用数学模型、统计学和计算机编程等技术手段来分析金融市场和制定投资策略的方法。通过量化金融，投资者可以利用大量的数据和算法来进行系统性的分析，以辅助决策和优化投资组合。这种方法通常涉及大规模的数据分析、模型构建和自动化交易执行。

其发展背景可以追溯到 20 世纪 60 年代和 70 年代，当时学者们开始运用数学和计量经济学方法来研究金融市场的行为，为量化金融金融积累了早期理论。20 世纪 80 年代，当时计算机技术的迅猛发展为大规模数据分析提供了可能，计算机参与到金融交易之中，是量化金融的早起实践。1987 年的黑色星期一股市崩盘成为量化金融发展的一个关键时刻。这场崩盘促使投资者和机构更加关注风险管理和模型化交易策略，引发了对量化方法在金融领域应用的兴趣。1990 年代，随着计算机性能的提升和金融市场数据的广泛数字化，

量化金融进入了一个新的阶段。崭新的算法和数学模型应运而生，为投资者提供了更准确和自动化的交易工具。同时，对市场数据的高频率交易（HFT）的兴起也推动了量化金融的快速发展 [1][2]。2007-2008 年的全球金融危机再次推动了量化金融的发展。危机爆发后，人们认识到传统的金融模型在应对极端市场情况上的局限性 [3]。因此，对于更复杂、全面的风险模型和交易策略的需求催生了更深入的量化金融研究 [4]。随着时间的推移，机器学习和人工智能技术的蓬勃发展为量化金融领域注入了新的活力。这些技术的广泛应用使得量化交易模型更加智能、灵活，并提高了对金融市场动态变化的适应能力 [5]。今天，量化金融已经成为金融领域中不可或缺的一部分，为投资者提供了更为精密和创新的交易工具。

2.2. 量化金融中的关键过程以及人工智能的参与

现在的量化金融的工作流程已经相对成熟并成体系化，其中主要涉及以下几个关键步骤，而人工智能均可应用于量化金融的这些环节之中，并在这些环节中发挥重要作用：

- **数据收集与清洗：**数据是量化金融的素材。量化金融的第一步就是获取市场数据、财务报表等金融相关信息，并进行数据清洗，确保数据的准确性和一致性。而人工智能技术可以用于大规模数据的自动收集、清洗和处理，提高数据处理的效率，同时通过算法识别异常值，确保数据的质量。
- **特征工程：**选择和构建合适的特征，以揭示市场中的潜在模式和趋势。机器学习算法能够自动提取关键特征，发现非线性关系，进一步优化特征工程的过程。
- **模型开发与训练：**基于已经收集的历史数据以及选定的特征，开发量化模型，并进行训练以使其适应市场的动态变化。人工智能技术，特别是卷积神经网络和随机森林算法，可以用于建立预测模型、时序模型等，可以通过学习大量历史数据来发现模式，提高模型的准确性。这一点也是量化金融区别于传统金融手段的重要的不同点。
- **回测与优化：**使用历史数据对开发的模型进行回测，评估其性能，并根据结果对模型进行优化。人工智能技术可用于自动化回测流程，同时通过我们接下来要介绍的遗传算法等优化算法寻找模型参数的最佳组合。

通过在这些关键步骤中整合人工智能技术，量化金融能够更智能地应对复杂多变的金融市场环境，提高决策的精度和效率。

3. 研究现状

3.1. 概述

之前的研究中验证了使用多种因子与某些机器学习算法相结合，在预测股票收益率方面有显著效能 [6][7]。本研究探讨了随机森林、卷积神经网络、配对交易和遗传算法在量化金融中的应用。通过数据回测实验，我们取得了令人满意的结果，显示了这些算法在股票市场预测方面的潜力和效能。这些研究结果为量化金融提供了新的视角和方法，有望改进投资和交易策略的决策基础。我们通过将算法的理论与本文的研究领域进行结合，在前人的基础上开发了有较好性能的实证模型，验证了算法的有效性，并且分析了部分算法的局限性。我们还尝试从理论上分析有效性和局限性的来源。本节将对这些算法本身和每一个算法在量化金融中的建模方法也进行详细介绍。

3.2. 卷积神经网络

3.2.1. 概述

卷积神经网络（Convolutional Neural Network，CNN）在计算机视觉领域的成功应用引发了对其在其他领域的研究兴趣 [8]，其中包括金融领域 [9]。卷积神经网络具有一些独特的特点和优势，这些特点在金融的量化分析领域尤其有价值，特别是在处理股票序列数据方面：

- **局部感知能力：**股价波动常受到局部事件和因素的影响，如公司公告和宏观经济数据。

CNN 的卷积操作有助于捕获这些局部特征，提高对短期波动的准确性。

- **参数共享：**股价数据通常非平稳，数据有限。CNN 通过参数共享降低参数数量，减少过拟合风险，特别适用于数据稀缺的金融领域。
- **平移不变性：**股价特征通常与时间相关，但不依赖于特定时间点。CNN 的平移不变性允许模型在不考虑时间点的情况下捕捉到这些特征，有助于识别趋势和模式。
- **层次性特征提取：**股价数据包含多个层次的信息，从短期波动到长期趋势。CNN 的多层结构逐渐提取不同层次的特征，例如第一层捕捉短期波动，后续层捕捉更高级的趋势，有助于理解复杂的股价结构。

3.2.2. 卷积神经网络在股票交易择时的应用

在过往的研究中，在限价订单簿（LOB）数据的基础上，已经有不少研究 [10][11] 在本研究中，我们构建了一个股票涨跌预测模型，旨在利用市场数据来预测股票的涨跌情况。我们将某支股票的开盘价、收盘价、市值、市盈率和最高价作为输入因子，并进行了标准化处理，以保证数据的一致性和可比性。我们将每 60 天的数据作为一个输入序列，然后预测下一个交易日的股票涨跌情况作为输出。

模型的核心架构是一个具有两层卷积神经网络（CNN）的深度学习模型。我们选择使用 CNN 来捕捉输入序列中的序列特征，因为 CNN 在处理序列数据方面具有出色的性能。每个卷积层后面跟着 ReLU 激活函数以引入非线性特性，并使用 Dropout 来减少过拟合的风险。接下来，我们通过全连接层（FC）将卷积层提取的特征组合在一起，然后经过 ReLU 激活函数进一步处理。最后的全连接层使用 softmax 函数输出一个包含两个元素的概率向量，表示股票涨跌的概率。

为了评估我们的模型性能，我们将数据集划分为三个部分：训练集、测试集和验证回测集。具体来说，我们选取了 2005 年至 2018 年的数据作为训练集，其中 20% 的数据用作测试集。而 2019 年的数据则用于验证回测。在回测中，我们根据模型的涨跌预测，在每个交易日采取相应的交易策略。如果模型预测股票涨，我们买入 1/3 的资金；如果模型预测股票跌，我们卖出 1/3 的仓位。如果已满仓或空仓，则不做任何买卖操作。

模型性能与实验细节将在第四部分实验中讲解。

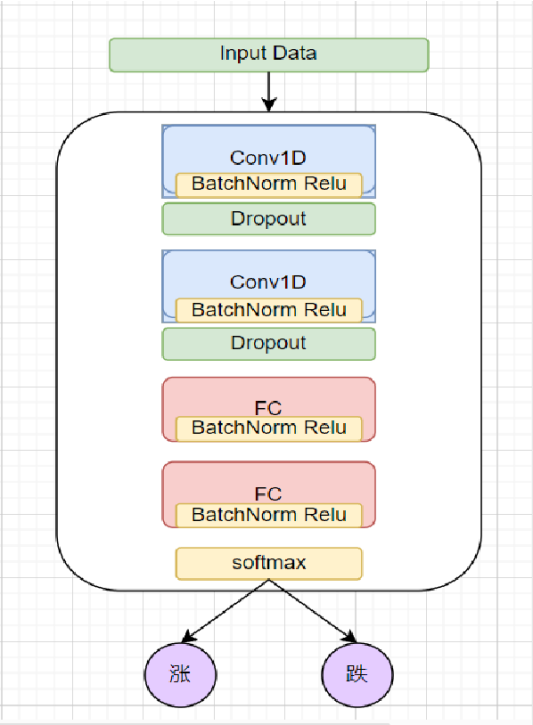

Figure 1: 卷积神经网络模型架构

3.3. 随机森林

3.3.1. 概述

随机森林是一种集成算法，它属于 Bagging 类型，通过组合多个弱学习器（决策树），对弱学习器的结果投票或取均值得到整体模型的最终结果，使得整体模型的结果具有较高的精确度和泛化性能。其之可以取得不错成绩，主要归功于“随机”和“森林”，一个使它具有抗过拟合能力，一个使它更加精准。

- **决策树**：决策树是一种流程图形式的模型，用于进行分类和回归分析。它通过树状图的结构来表示在一个决策过程中的各种可能的结果，以及导致这些结果的各种决策路径。日

常生活中，我们对于事物的认知都是基于特征的判断与分类，譬如通过胎生与否可判断哺乳动物，根据肚脐尖圆来挑选螃蟹公母。决策树就是采用这样的思想，基于多个特征进行分类决策。在树的每个结点处，根据特征的表现通过某种规则分裂出下一层的叶子节点，终端的叶子节点即为最终的分类结果。决策树学习的关键是选择最优划分属性。随着逐层划分，决策树分支结点所包含的样本类别会逐渐趋于一致，从而得到最终分类。

- **基尼不纯度**：决策树的分裂遵循基尼不纯度的最大减少。节点的基尼不纯度是指，根据节点中样本的分布对样本分类时，从节点中随机选择的样本被分错的概率。因此，基尼不纯度越小，样本的纯度越高。
- **自助采样**：随机森林在构建每棵决策树时采用自助采样方法。对于每棵树而言，从原始训练数据中有放回地随机抽取样本，形成一个新的训练数据集。这样就使得每棵决策树的训练数据都略有不同，从而增加了模型的多样性。
- **多数投票机制**：在随机森林中，对于分类问题，每棵决策树都投票给某个类别，最终的预测结果是得票最多的类别。这种投票机制对单棵决策树可能存在的错误做出了整体性的修正，从而提高了模型的泛化能力。

3.3.2. 随机森林在选股中的应用

几十年来，对股票回报主要驱动因素的调查一直备受关注。正如经典的金融理论所示，股票回报以线性方式归因于基础基本面，包括系统风险、市值、市净率等。相应的线性回归方案，结合广泛发展的涵盖各种技术和基本方面的因子，构成学术和行业领域金融建模的主要工具。股票选择在投资组合管理中扮演着重要角色。股票选择标准是复杂的，绝不符合科学标准。然而，新开发的先进算法，如深度神经网络和基于树的模型，相对于传统的因子加权排名系统，为股票选择效率提供了新视角，从而导致显著的交易业绩。

决策树可用于各种机器学习应用。但是，为了学习高度不规则的模式，生长得非常深的树往往会过拟合训练集。数据中的轻微噪音可能导致树以完全不同的方式生长。这是因为决策树具有很低的偏差和很高的方差。随机森林通过在特征空间的不同子空间上训练多个决策树来克服这个问题，代价是稍微增加了偏差。这意味着森林中的任何一棵树都没有看到整个训练数据。数据被递归地分割成分区。在特定节点，通过对属性提出问题来进行拆分。选择拆分标准是基于一些不纯度度量，如香农熵或基尼不纯度。

我们调研了一些包含实验的随机森林选股论文，从他们的实验结论中获取合适的参数。Zheng Tan 等人的工作 [12] 指出：树的数量对结果有显著影响当树的数量设置为 60 时，夏普比率可达 2.75，同样，Sortino 和 Calmar 比率也能达到它们的最大值。样本内 oob 分数随着树的数量稳步增加，但随着树的数量接近 120，趋势趋于平稳。此外，随着样本类别数量的增加，样本外绩效投资组合和对冲的 NV 都有所恶化，尤其是在 2016 年和 2017 年，更多的样本类别导致超额回报减少。另一篇是 Khaidem L[13] 等人的工作，他们的森林有 30 颗树，对比了 30,60,90 天的交易周期，发现 90 天的效果最好。尽管我们的随机森林建模是通过 QuantConnect 平台的可视化平台实现的，也就是说代码的实现细节可能和这两篇的模型有出入，但调参时我们充分参考了两篇论文实验部分的参数和结果。

3.4. 配对算法

3.4.1. 核心思想

配对交易 [14, 15] 是一种基于统计、机器学习或深度学习的套利的策略，其核心在于利用两种证券价格之间的历史相关性进行套利。这种策略的基本假设是，如果两种证券在过去的某段时间内价格走势紧密相关，那么它们未来的价格走势也将保持这种相关性。当这两种证券的价格差异超出历史正常范围时，投资者会同时买入表现较弱的证券并卖空表现较强的证券，期望未来这两

种证券的价格差异会回归到历史正常水平。这种策略的优势在于其市场中性特性，即它的盈利不受市场整体涨跌的影响。

在本段中，我们将会给出传统的基于统计量的经典算法 [14]，以及采用机器学习、利用多种聚类算法辅助的新方案 [15]。

3.4.2. 基于统计的配对交易算法

在 Gatev 的工作中 [14]，配对交易的实施主要基于历史价格数据的最小距离匹配。具体算法如下：

- **股票对的选择**：通过计算股票对之间的历史价格归一化后的欧几里得距离，选择距离最小的股票对作为交易对象。这种方法的核心在于识别那些历史上价格走势紧密相关的股票对，并利用这种相关性来预测未来的价格走势 [14]。

$$\text{距离} = \sqrt{\sum_{t=1}^{T}(P_{1,t} - P_{2,t})^2}$$

  其中，$P_{1,t}$ 和 $P_{2,t}$ 分别代表两只股票在时间 t 的价格。
- **交易信号的触发**：当选定的股票对价格差异超过设定的阈值时，触发交易信号。通常这个阈值是基于历史价格差异的标准差来设定的。
- **交易执行**：在价格差异超过阈值时，投资者会买入表现较弱的股票并卖空表现较强的股票，等待价格差异回归正常。

此外，该工作还探讨了配对交易盈利的稳定性，论文指出配对交易的盈利可能与执行此策略的市场参与者的交易成本和做空能力有关。

3.4.3. 基于无监督学习的配对交易算法

另一个工作中 [15]，提出了一种新的配对交易策略，该策略利用无监督学习来识别股票对。与传统的基于时间序列回报的配对交易策略不同，这种方法结合了公司特征和价格信息来识别配对交易的机会。

核心思想：传统的配对交易策略主要基于股票的历史价格数据，而这项研究通过结合公司的财务特征和市场表现等信息，使用无监督学习方法来识别具有相似特征的股票对。这种方法的创新之处在于，它不仅仅依赖于价格数据，而是结合了更多维度的信息来发现价格走势上的相关性。

无监督学习算法选择：论文中采用了三种代表性的聚类方法：k-means 聚类 [16, 17]、DBSCAN[18] 和层次聚类 [19]。这些方法通过将股票基于过去的回报和特征进行分组，从而识别出潜在的配对交易机会。

- **k-means 聚类**：k-means 聚类 [16, 17] 通过最小化簇内平方和（WCSS）来将数据点分配到不同的簇中。

$$\text{WCSS} = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}||x_i - \mu_j||^2$$

  其中，x_i 是数据点，μ_j 是簇 j 的中心，z_{ij} 是指示变量，如果数据点 i 属于簇 j 则为 1，否则为 0。
- **DBSCAN 聚类 [18]**：DBSCAN 聚类基于高密度区域的连续性来识别簇，将低密度区域的点视为噪声。
- **层次聚类 [19]**：层次聚类是一种自底向上的方法，初始将每个数据点视为一个单独的簇，然后逐步合并最接近的簇。

$$d(A, B) = \min(||a-b||) \quad \text{对于所有 } a \in A, b \in B$$

  其中，$d(A, B)$ 是集群 A 和 B 之间的距离。

实验方案及结果：在论文实验中 [15]，当过去一个月的回报差异超出一个横截面标准差时，执行买入表现较弱股票和卖空表现较强股票的策略。应用于 1980 年至 2020 年的美国股市，通过层次聚类构建的多空组合实现了年化平均回报率 24.8% 和夏普比率 2.69 的显著成绩。因此，论文实验结果表明，引入无监督学习方案能显著提高了配对交易策略的表现。

3.5. 遗传算法

3.5.1. 概述

遗传算法是基于达尔文《进化论》中的遗传以及自然选择现象之上的一种搜索算法。它最早由 Holland 于 1975 年提出 [20]。遗传算法将问题的解视为个体，使用选择、交叉和变异三个遗传算子模拟生物的遗传和变异现象。算法通过自然选择淘汰不适应环境的个体最终得到最适应环境的个体，即问题的最优解。

遗传算法只要求目标函数是可计算的，对问题中约束、变量的数量、形式没有限制。因此，基于遗传算法的量化交易策略较为灵活，形式多样 [21]。在本节中，我们将介绍两种不同的基于遗传算法的交易策略。

3.5.2. 使用遗传算法生成交易策略

Hirabayashi 提出了一种自动生成交易策略的方法 [22]。该方法使用遗传算法寻找技术分析指标 [23] 的最佳参数以及组合方式，以此得到进行买入或卖出操作的最佳时机。方法的主要流程如下：

- 预先计算技术指标：预先计算在遗传算法运行过程中需要用到的各技术指标的值。文章中选择的技术指标为 RSI1，RSI2，PD 以及 RR。
- 对个体进行编码：为了方便遗传算法的交叉和突变操作，将一个个体使用一个二进制串表示。每个二进制串主要分为买策略、卖策略以及策略类型三个部分。其中买策略、卖策略又由各技术指标参数的值，参数的位置，以及技术指标之间的关系等多个部分组成。
- 遗传算法搜索：使用历史数据进行模拟交易，将策略的收益作为个体的适应度。使用联赛选择算法淘汰适应度较低的个体。经过多次遗传、选择，得到最优的交易策略。

作者使用 2005-2008 年之间的美元/日元，欧元/日元以及澳元/日元的外汇数据测试得到的策略的效果。实验显示自动生成的策略在固定的统计时间窗口中得到了正回报。

此方法使用遗传算法自动生成了有效的量化交易策略。然而，由于使用的技术指标种类，技术指标的参数范围等仍需要人工事先指定，不能做到完全的策略自动生成。此外，由于此方法生成的策略没有选股功能，仅能对单只股票/外汇进行分析，故此方法生成策略的效果仍会受选择的股票的走势所影响。

3.5.3. 使用遗传算法选择投资组合

针对单股票策略收益受股票波动影响大的问题，Matsumura 提出了一种使用遗传算法选择投资股票的方法 [24]。该方法在使用遗传算法生成策略之外，还加入了使用遗传算法生成最优投资组合的功能。方法主要由两“层”搜索算法组成：外层使用遗传算法寻找最优的投资组合，内层使用遗传编程 [25] 以及决策树算法，计算外层确定的投资组合的收益情况。其中，内层算法的主要流程与针对单只股票的策略生成方法相似，在此不过多赘述。外层算法主要流程如下：

- 确定待选股票组：首先需要指定一个包含较多股票的待选股票组，以便遗传算法能找到较优的投资组合。
- 对投资组合进行编码：投资组合主要包含两部分：选取的股票种类以及每种股票的投资比例。其中投资比例为一个介于 0 与 1 之间的实数，表示投入某种股票的资金占全部资金的比例。例如，若本金为 10 万元，股票 A 的投资比例为 0.2，则投入股票 A 的资金为 2 万元。由于需要保证所有股票的投资比例之和为 1，需要对投资比例进行正则化处理 [26]。
- 遗传算法搜索：对于每个投资组合，调用内层搜索算法获得其收益情况。将投资组合的收益情况视为个体的适应度，使用遗传算法生成最优的投资组合。

作者从日经 225 中选取了 56 只待选股票。根据 1997-2008 年的交易数据，使用上述方法生成了一个包含 10 只股票的交易策略。模拟交易显

示，生成的策略的收益率相较“买入并持有”策略收益更高，且该策略风险性较低。

4. 实验

4.1. 卷积神经网络

4.1.1. 实验的数据与参数

模型的核心架构是一个具有两层卷积神经网络（CNN）的深度学习模型。我们选择使用 CNN 来捕捉输入序列中的序列特征，因为 CNN 在处理序列数据方面具有出色的性能。每个卷积层后面跟着 ReLU 激活函数以引入非线性特性，并使用 Dropout 来减少过拟合的风险。接下来，我们通过全连接层（FC）将卷积层提取的特征组合在一起，然后经过 ReLU 激活函数进一步处理。最后的全连接层使用 softmax 函数输出一个包含两个元素的概率向量，表示股票涨跌的概率。模型架构图在第三部分已经有所讲解。

我们的实验数据使用的是 A 股的股票，年份是 2005-2018。本实验将某支股票的 [开盘、收盘、市值、市盈率、最高价] 作为因子输入，然后标准化处理。每连续 60 天的数据作为输入，涨跌预测为输出，以滚动方式循环数据。涨跌的标签由这 60 天之后的那一个交易日与前一日的变化作为标签。当后一天股价收盘价比前一天高，则标签为涨，反之则为跌。

数据集分为训练集、测试集、验证回测集。我们取 2005-2018 年为训练集，其中抽取 20% 为测试集。将 2019 年数据作为验证回测集合。初始化股票仓位为空仓，当预测下一天趋势为涨，当日买入 1/3 金额，若预测为跌，则卖出 1/3 金额的仓位。若满仓、空仓则分别不进行买入、卖出操作。

本实验使用的是单只股票的交易数据五元组，我们分别测试了贵州茅台、工商银行、比亚迪三只股票。下面实验数据以贵州茅台举例。

4.1.2. 实验的训练与结果

实验共进行了 200 个 epoch，初始学习率为 0.01，每进行 5 个 epoch 学习率变为原来的 0.9，优化器为 SGD。在每一个循环中统计其训练误差（蓝色线）、测试误差（黄色线）、测试准确率，并且画出图如 figure 2。下面是关于 figure 2的一些解释：

- Training Loss 波动但大趋势下降：波动可能表明学习率设置不太稳定或数据集中存在一些不规则特征。
- Test Loss 先升高后下降：初始升高可能表明模型在最初阶段对测试数据过拟合或学习率较高。学习率下降之后效果逐渐变好。
- Test Accuracy 先不变后缓慢升高：初始学习率过高或者特征难以把握，随后的缓慢提升表明模型逐渐开始泛化。最终 accuracy 稳定在 0.54 左右波动。

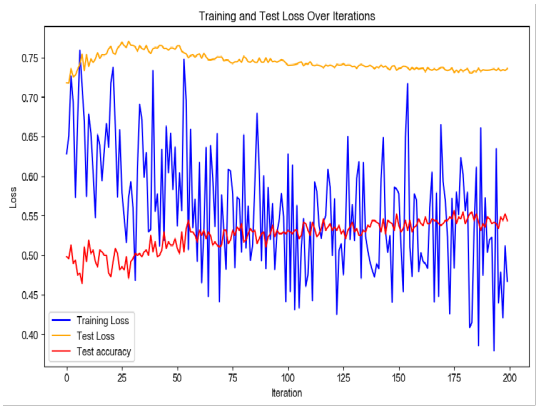

Figure 2: Loss and accuracy of Training

我们通过记录下训练过程中 accuracy 最高的模型，然后使用它来进行验证回测。输入前 60 天的 [开盘、收盘、市值、市盈率、最高价] 五元组作为特征，当预测下一天趋势为涨，当日买入 1/3 金额，若预测为跌，则卖出 1/3 金额的仓位。若满仓、空仓则分别不进行买入、卖出操作。本实验中使用 BigQuant 作为回测平台。其中使用茅台的数据回测时，回测结果如 figure3所示。红色线为回测结果，蓝色线为基准收益率。

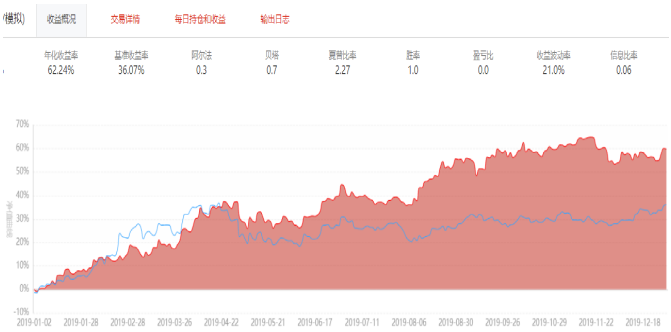

Figure 3: Results of model backtesting

下面是本实验的一些关键结果的数值

回测收益-茅台	
结果参数	数值
年化收益率	62.24%
基准收益率	36.07%
夏普比率	2.27
收益波动率	21.0%

Table 1: 以茅台为例的实验数据

根据上述结果可以看到，模型能够在过去取得不错的效果。不过由于本实验的模型较为简单，对于不同类型的股票性能不尽相同。且模型对于不同规律的捕捉较为迟钝，比如疫情到来之后，模型对于变化之后的环境，适应度不足。具体的针对卷积神经网络的即时改进，还有待下一步探索。

4.2. 随机森林

4.2.1. 实验环境与参数

使用 QuantConnect 平台进行随机森林算法建模和测试与验证 [27]。QuantConnect 是一个开源、云基础的算法交易平台，为用户提供股票、外汇、期货、期权、衍生品和加密货币等多种市场的算法交易服务。我们使用该平台的可视化策略来完成建模。下面是代码架构

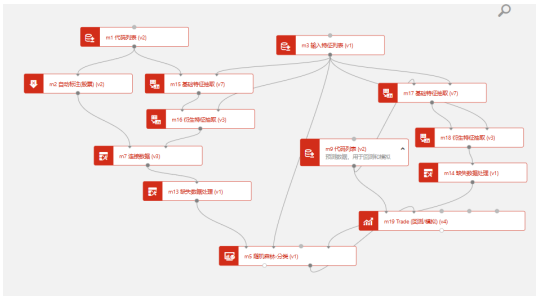

Figure 4: 随机森林代码架构

代码中比较重要的是 m19。m19 的 handle data 方法用于每日数据处理函数，每天执行一次。先按日期过滤得到今日的预测数据。然后是资金分配，平均持仓时间是 hold days（后面简称 h），每日都将买入股票，每日预期使用 1/h 的资金。实际操作中，会存在一定的买入误差，所以在前 h 天，等量使用资金；之后，尽量使用剩余资金。接着生成卖出订单，h 天之后才开始卖出；对持仓的股票，按机器学习算法预测的排序末位淘汰。最后生成买入订单，按机器学习算法预测的排序，买入前面一定数量股票，需要确保股票持仓量不会超过每次股票最大的占用资金量。initialize 方法是初始化函数。先加载预测数据，设置交易手续费和滑点。接着设置买入的股票数量，这里默认买入预测股票列表排名靠前的 5 只，不过后来经过不断调参，发现这里设成 20 以上后效果明显更好，35 的时候最好。最后设置的股票的权重，该权重分配会使得靠前的股票分配多一点的资金。

4.2.2. 实验结果

获取 2010-2015 年 A 股交易数据进行学习，预测 2015-2016 的股票。经过不断调参，随机森林的参数选择如下：

树的数量	60
特征使用率	0.8
树的最大深度	5
每叶子节点最小样本数	500

我们尝试改变参数来优化回测结果，根据调研的论文提供的信息，决策树的个数越多，模型越复杂，计算速度越慢；每棵树的最大深度大则

拟合能力强，数值小则泛化能力强；每个叶子节点最少样本数：数值大泛化能力强，数值小拟合能力强。但是经过实验发现，改变这些对实际决策的影响都很小，反而是买入的股票数量对回测效果影响很大。默认为 5 的时候效果很差，到了 20 以上会比较好。下面是 35 时候，也就是最好情况的回测数据：

Figure 5: 随机森林算法收益率

运行天数	365 天
最大回撤	35.83%
年化收益率	74.52%
夏普比率	1.59
信息比率	0.18

参考论文中的最好结果是 2.25 的夏普比率和 84.96% 的年回报率，可以看出我们的模型与论文中仍有差距。考虑到决策树数量、树最大深度、每叶节点等与过拟合相关的参数调整后模型的决策变化不大，我们认为与论文的差距是是模型的实现细节导致的。但整体而言，最终回测数据还是比较理想，说明随机森林算法用于选股有不错的效果。

4.3. 配对算法

4.3.1. 实验环境

本研究采用 QuantConnect 平台进行算法交易的测试与验证 [27]。QuantConnect 是一个开源、云基础的算法交易平台，为用户提供股票、外汇、期货、期权、衍生品和加密货币等多种市场的算法交易服务。该平台支持 Python 和 C# 等多种编程语言，并通过其 LEAN 算法交易引擎实现算法的设计、回测和交易 [28]。

在 QuantConnect 平台 [27] 上，我们实验了配对交易策略 [14]，该策略基于统计套利理念，寻找并交易两个价格序列之间的长期关系。交易算法主要通过计算两只 ETF 之间的协整关系和相关性，来识别和利用资产价格的微小偏差。在策略构建中，实验选择了多个 ETF 作为潜在的交易对，并通过 Pearson 相关系数模型 [29] 和协整向量模型 [30] 来选择最佳配对并构建投资组合。

4.3.2. 实验方法

配对交易算法通过 QuantConnect 平台实施，选取包括 IYM、XLE、XLK、XLF 和 XLI 在内的 ETF 作为可交易对象。算法启动日期为 2020 年 4 月 8 日，初始资金设定为 100 万美元。在实现上，设置回溯窗口 60 天来计算相关性和协整关系，预期利润率至少为 2% 以覆盖交易费用。此外，算法使用分钟级分辨率数据，并通过滚动窗口和日志收益指标来分析和预测各 ETF 之间的协整向量 [30]。

Pearson 相关系数和协整关系。Pearson 相关系数度量两个变量之间的线性相关性，公式为 [29]：

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

协整关系则用于检测不同时间序列之间的长期稳定关系。Engle-Granger 协整测试是识别协整关系的一种方法 [**coint**]，通过以下回归模型进行：

$$Y_t = \alpha + \beta X_t + \epsilon_t$$

其中，如果残差序列 ϵ_t 是平稳的，则 Y_t 和 X_t 被认为是协整的。

4.3.3. 实验结果

策略在 1293 个交易日内运行，期间最大回撤为 14.4%。交易周转率为 11%，显示策略具有适度的交易活动。年化复合增长率（CAGR）为 1.7%，显示了策略在测试期间的增长能力。而夏普比率为 0.0，可能表明策略的收益与无风险收益相当。信息比率为 0.2，表明策略相对于基准的超额回报。策略容量估计为 240 万美元。

运行天数	1293 天
最大回撤	14.4%
周转率	11%
年化复合增长率（CAGR）	1.7%
夏普比率	0.0
信息比率	0.2
每日交易次数	0.2
策略容量（美元）	240 万

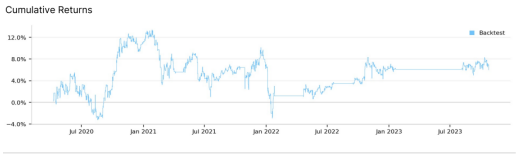

Figure 6: 配对算法收益率

4.4. 遗传算法

4.4.1. 实验环境与方法

我们在 BigQuant 平台实现了基于 Hirabayashi 提出的生成单股交易策略算法的模型 [22]。该模型选用的技术指标为 MACD 与 KDJ，每个个体使用 34 个比特编码技术指标的参数及指标之间的关系。在计算个体适应性的部分，我们根据平台提供的股票历史数据，通过模拟交易得到个体的收益率，并将其作为个体的适应性。模拟交易的主要流程如下：

- 初始化：由于部分技术指标使用股票历史股价的加权平均值对股票的趋势进行分析预测，如 MACD 指标使用的是股票前 34 个交易日的股价指数加权平均。所以需要在进行交易前使用股票的历史股价信息进行初始化。
- 交易：在每个交易日结束后，更新技术指标，并根据策略发出的买卖信号进行操作：若当前为空仓且策略发出买入信号，则在第二天开盘时满仓对应股票；若当前满仓且策略发出卖出信号，则在第二天开盘时平仓。
- 收益率计算：在模拟交易结束后，使用如下公式计算收益率：

$$returns = \frac{cash + stock}{initial} - 1$$

其中，$initial$ 为初始资金，$cash$ 为交易结束后的资金量，$stock$ 为交易结束后股票市值。

同时，我们使用 BigQuant 平台提供的回测功能测试生成的策略的收益率，其中交易流程与上述遗传算法中模拟交易流程相同。

4.4.2. 实验数据与参数

我们使用南方航空公司 2017 年 1 月 1 日至 2021 年 1 月 1 日的股价信息作为训练集，2021 年 1 月 1 日至 2022 年 1 月 1 日之间的股价信息作为验证集，进行策略的生成及验证。其它的参数如下：

个体数量	100
迭代次数	25
交叉互换概率	0.1
突变概率	0.7

4.4.3. 实验结果

Figure 7: 遗传算法回测报告

在 2021 年 1 月 1 日至 2022 年 1 月 1 日之间，“买入并持有”策略的收益率为 15.42%，而生成的策略的收益率为 29.35%，高于“买入并持有策略”13.93%。夏普比率为 1.13，表明生成的策略有不错的表现。

然而，策略的最大回撤 13.65%，收益波动率为 23.49%，仍有优化的空间。观察回测报告发现，生成的策略在交易过程中满仓时间占比较大，导致策略的收益受股价波动影响较大，这可能是策略最大回撤以及收益波动率较大的原因之一。

5. 结论

我们小组以卷积神经网络、随机森林、配对算法与遗传算法为代表，研究了人工智能在量化

金融领域中的应用，分别对其进行了模拟实验与数据回测且效果令人满意，展示了人工智能在处理量化金融领域有关问题的优越性。

在卷积神经网络的数据回测中，我们以贵州茅台股票为例，发现在卷积神经网络的应用下取得了相较于基准收益而言非常可观的收益率，并把收益波动控制在了合理范围。但也在实践中发现了模型对于不同规律的捕捉较为迟钝，比如疫情到来之后，模型对于变化之后的环境，适应度不足。卷积神经网络的即时改进依然有待优化。在使用 QuantConnect 平台对随机森林算法建模测试的过程中，我们实现了 74.52% 的年化收益率，虽然由于实现细节方面的差异，与参考论文中的 84.96% 仍有一定差距，但是总体效果也令人满意，体现了随机森林算法对选股的指导作用。同时我们还发现了决策树数量、树最大深度与每叶节点等于过拟合相关的参数调整后对决策在一定程度上总体影响不大。在配对算法的研究中，我们让选定的配对策略在 1293 个交易日模拟运行，取得了比基准收益明显更高的收益，提现了配对算法的优势。最后在遗传算法的模型中，我们的收益率和夏普比率同样具有优势，同时发现了模型在最大回撤与收益波动方面仍有优化空间。

总体而言，在我们的模拟实验与数据回测之中，模型在收益率与夏普比率方面的优势明显，提供了具有参考价值的实践成果与实验数据。同时，我们通过模拟实验也发现了领域内现有理论模型在处理具体问题时的一些局限性与不足，对量化金融的决策与优化给出了合理建议与改进方向，对于人工智能参与下的量化金融决策及其未来的优化与改良都有借鉴意义。

Reference

[1] Tse, J.; Lin, X.; Vincent, D. High frequency trading—Measurement, detection and response. *Credit Suisse, Zürich, Switzerland, Tech. Rep* **2012**.

[2] Avellaneda, M. In *New York University & Finance Concepts LLC, Quant Congress USA*, **2011**.

[3] Neuhierl, A.; Varneskov, R. T. Frequency dependent risk. *Journal of Financial Economics* **2021**, *140*, 644–675.

[4] Riordan, R.; Storkenmaier, A. Latency, liquidity and price discovery. *Journal of Financial Markets* **2012**, *15*, 416–437.

[5] Ntakaris, A.; Magris, M.; Kanniainen, J.; Gabbouj, M.; Iosifidis, A. Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods. *Journal of Forecasting* **2018**, *37*, 852–866.

[6] Bollerslev, T.; Marrone, J.; Xu, L.; Zhou, H. Stock Return Predictability and Variance Risk Premia: Statistical Inference and International Evidence. *Journal of Financial and Quantitative Analysis* **2014**, *49*, 633–661.

[7] Ferreira, M. A.; Santa-Clara, P. Forecasting stock market returns: The sum of the parts is more than the whole. *Journal of Financial Economics* **2011**, *100*, 514–537.

[8] Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **1998**, *86*, 2278–2324.

[9] Sangadiev, A.; Rivera-Castro, R.; Stepanov, K.; Poddubny, A.; Bubenchikov, K.; Bekezin, N.; Pilyugina, P.; Burnaev, E. DeepFolio: Convolutional neural networks for portfolios with limit order book data. *arXiv preprint arXiv:2008.12152* **2020**.

[10] Zhang, Z.; Zohren, S.; Roberts, S. DeepLOB: Deep Convolutional Neural

Networks for Limit Order Books. *IEEE Transactions on Signal Processing* **2019**, *67*, 3001–3012.

[11] Parlour, C. A.; Seppi, D. J. Limit order markets: A survey. *Handbook of financial intermediation and banking* **2008**, *5*, 63–95.

[12] Tan, Z.; Yan, Z.; Zhu, G. Stock Selection with Random Forest: An Exploitation of Excess Return in the Chinese Stock Market. *Heliyon* **2019**, *5*, e02347.

[13] Khaidem, L.; Saha, S.; Dey, S. R. Predicting the Direction of Stock Market Prices Using Random Forest. *arXiv preprint* **2016**.

[14] Gatev, E.; Goetzmann, W. N.; Rouwenhorst, K. G. Pairs Trading: Performance of a Relative Value Arbitrage Rule, Rochester, NY, **2006**.

[15] Han, C.; He, Z.; Toh, A. J. W. Pairs Trading via Unsupervised Learning, Rochester, NY, **2021**.

[16] MacQueen, J. In **1967**.

[17] In *Wikipedia*, Page Version ID: 1186084028, **2023**.

[18] Ester, M.; Kriegel, H.; Sander, J.; Xu, X. In Knowledge Discovery and Data Mining, **1996**.

[19] Johnson, S. C. Hierarchical clustering schemes. *Psychometrika* **1967**, *32*, 241–254.

[20] Holland, J. H., Adaptation In Natural and Artificial Systems; MIT Press: **1975**.

[21] Aguilar-Rivera, R.; Valenzuela-Rendón, M.; Rodríguez-Ortiz, J. Genetic algorithms and Darwinian approaches in financial applications: A survey. *Expert Systems with Applications* **2015**, *42*, 7684–7697.

[22] Hirabayashi, A.; Aranha, C.; Iba, H. In *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation*, Association for Computing Machinery: Montreal, Québec, Canada, **2009**, 1529–1536.

[23] Murphy, J., Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications; New York Institute of Finance Series; Penguin Publishing Group: **1999**.

[24] Matsumura, K.; Kakinoki, H. Portfolio Strategy Optimizing Model for Risk Management Utilizing Evolutionary Computation. *Electronics and Communications in Japan* **2014**, *97*, 45–62.

[25] Koza, J. R., Genetic Programming: On the Programming of Computers by Means of Natural Selection; MIT Press: Cambridge, MA, USA, **1992**.

[26] Xia, Y.; Liu, B.; Wang, S.; Lai, K. A model for portfolio selection with order of expected returns. *Computers and Operations Research* **2000**, *27*, Cited by: 161, 409–422.

[27] Design and trade algorithmic trading strategies in a web browser, with free financial data, cloud backtesting and capital - QuantConnect.com `https://www.quantconnect.com/` (accessed 12/27/2023).

[28] In *Wikipedia*, Page Version ID: 1158099323, **2023**.

[29] Pearson correlation coefficient - Wikipedia `https://en.wikipedia.org/wiki/Pearson_correlation_coefficient` (accessed 12/27/2023).

[30] Engle, R. F.; Granger, C. W. J. Co-Integration and Error Correction: Representation, Estimation, and Testing. *Econometrica* **1987**, *55*, 251.